Graduate Texts in Mathematics

Rajendra Bhatia

Matrix Analysis

Springer

Graduate Texts in Mathematics **169**

Springer-Science+Business Media, LLC

Graduate Texts in Mathematics

1 TAKEUTI/ZARING. Introduction to Axiomatic Set Theory. 2nd ed.
2 OXTOBY. Measure and Category. 2nd ed.
3 SCHAEFER. Topological Vector Spaces.
4 HILTON/STAMMBACH. A Course in Homological Algebra.
5 MAC LANE. Categories for the Working Mathematician.
6 HUGHES/PIPER. Projective Planes.
7 SERRE. A Course in Arithmetic.
8 TAKEUTI/ZARING. Axiomatic Set Theory.
9 HUMPHREYS. Introduction to Lie Algebras and Representation Theory.
10 COHEN. A Course in Simple Homotopy Theory.
11 CONWAY. Functions of One Complex Variable I. 2nd ed.
12 BEALS. Advanced Mathematical Analysis.
13 ANDERSON/FULLER. Rings and Categories of Modules. 2nd ed.
14 GOLUBITSKY/GUILLEMIN. Stable Mappings and Their Singularities.
15 BERBERIAN. Lectures in Functional Analysis and Operator Theory.
16 WINTER. The Structure of Fields.
17 ROSENBLATT. Random Processes. 2nd ed.
18 HALMOS. Measure Theory.
19 HALMOS. A Hilbert Space Problem Book. 2nd ed.
20 HUSEMOLLER. Fibre Bundles. 3rd ed.
21 HUMPHREYS. Linear Algebraic Groups.
22 BARNES/MACK. An Algebraic Introduction to Mathematical Logic.
23 GREUB. Linear Algebra. 4th ed.
24 HOLMES. Geometric Functional Analysis and Its Applications.
25 HEWITT/STROMBERG. Real and Abstract Analysis.
26 MANES. Algebraic Theories.
27 KELLEY. General Topology.
28 ZARISKI/SAMUEL. Commutative Algebra. Vol.I.
29 ZARISKI/SAMUEL. Commutative Algebra. Vol.II.
30 JACOBSON. Lectures in Abstract Algebra I. Basic Concepts.
31 JACOBSON. Lectures in Abstract Algebra II. Linear Algebra.
32 JACOBSON. Lectures in Abstract Algebra III. Theory of Fields and Galois Theory.
33 HIRSCH. Differential Topology.
34 SPITZER. Principles of Random Walk. 2nd ed.
35 WERMER. Banach Algebras and Several Complex Variables. 2nd ed.
36 KELLEY/NAMIOKA et al. Linear Topological Spaces.
37 MONK. Mathematical Logic.
38 GRAUERT/FRITZSCHE. Several Complex Variables.
39 ARVESON. An Invitation to C^*-Algebras.
40 KEMENY/SNELL/KNAPP. Denumerable Markov Chains. 2nd ed.
41 APOSTOL. Modular Functions and Dirichlet Series in Number Theory. 2nd ed.
42 SERRE. Linear Representations of Finite Groups.
43 GILLMAN/JERISON. Rings of Continuous Functions.
44 KENDIG. Elementary Algebraic Geometry.
45 LOÈVE. Probability Theory I. 4th ed.
46 LOÈVE. Probability Theory II. 4th ed.
47 MOISE. Geometric Topology in Dimensions 2 and 3.
48 SACHS/WU. General Relativity for Mathematicians.
49 GRUENBERG/WEIR. Linear Geometry. 2nd ed.
50 EDWARDS. Fermat's Last Theorem.
51 KLINGENBERG. A Course in Differential Geometry.
52 HARTSHORNE. Algebraic Geometry.
53 MANIN. A Course in Mathematical Logic.
54 GRAVER/WATKINS. Combinatorics with Emphasis on the Theory of Graphs.
55 BROWN/PEARCY. Introduction to Operator Theory I: Elements of Functional Analysis.
56 MASSEY. Algebraic Topology: An Introduction.
57 CROWELL/FOX. Introduction to Knot Theory.
58 KOBLITZ. p-adic Numbers, p-adic Analysis, and Zeta-Functions. 2nd ed.
59 LANG. Cyclotomic Fields.
60 ARNOLD. Mathematical Methods in Classical Mechanics. 2nd ed.

continued after index

Rajendra Bhatia

Matrix Analysis

Springer

Rajendra Bhatia
Indian Statistical Institute
New Delhi 110 016
India

Mathematics Subject Classification (1991): 15-01, 15A16, 15A45, 47A55, 65F15

Library of Congress Cataloging-in-Publication Data
Bhatia, Rajendra, 1952–
Matrix analysis / Rajendra Bhatia.
p. cm. – (Graduate texts in mathematics ; 169)
Includes bibliographical references and index.
ISBN 978-1-4612-6857-4 ISBN 978-1-4612-0653-8 (eBook)
DOI 10.1007/978-1-4612-0653-8
1. Matrices. I. Title. II. Series.
QA188.B485 1996
512.9′434–dc20 96-32217

Printed on acid-free paper.

Production managed by Victoria Evarretta; manufacturing supervised by Jeffrey Taub.
Photocomposed pages prepared from the author's LaTeX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-6857-4 SPIN 10524488

Preface

A good part of matrix theory is functional analytic in spirit. This statement can be turned around. There are many problems in operator theory, where most of the complexities and subtleties are present in the finite-dimensional case. My purpose in writing this book is to present a systematic treatment of methods that are useful in the study of such problems.

This book is intended for use as a text for upper division and graduate courses. Courses based on parts of the material have been given by me at the Indian Statistical Institute and at the University of Toronto (in collaboration with Chandler Davis). The book should also be useful as a reference for research workers in linear algebra, operator theory, mathematical physics and numerical analysis.

A possible subtitle of this book could be *Matrix Inequalities*. A reader who works through the book should expect to become proficient in the art of deriving such inequalities. Other authors have compared this art to that of cutting diamonds. One first has to acquire hard tools and then learn how to use them delicately.

The reader is expected to be very thoroughly familiar with basic linear algebra. The standard texts *Finite-Dimensional Vector Spaces* by P.R. Halmos and *Linear Algebra* by K. Hoffman and R. Kunze provide adequate preparation for this. In addition, a basic knowledge of functional analysis, complex analysis and differential geometry is necessary. The usual first courses in these subjects cover all that is used in this book.

The book is divided, conceptually, into three parts. The first five chapters contain topics that are basic to much of the subject. (Of these, Chapter 5 is more advanced and also more special.) Chapters 6 to 8 are devoted to

perturbation of spectra, a topic of much importance in numerical analysis, physics and engineering. The last two chapters contain inequalities and perturbation bounds for other matrix functions. These too have been of broad interest in several areas.

In Chapter 1, I have given a very brief and rapid review of some basic topics. The aim is not to provide a crash course but to remind the reader of some important ideas and theorems and to set up the notations that are used in the rest of the book. The emphasis, the viewpoint, and some proofs may be different from what the reader has seen earlier. Special attention is given to multilinear algebra; and inequalities for matrices and matrix functions are introduced rather early. After the first chapter, the exposition proceeds at a much more leisurely pace. The contents of each chapter have been summarised in its first paragraph.

The book can be used for a variety of graduate courses. Chapters 1 to 4 should be included in any course on Matrix Analysis. After this, if perturbation theory of spectra is to be emphasized, the instructor can go on to Chapters 6,7 and 8. With a judicious choice of topics from these chapters, she can design a one-semester course. For example, Chapters 7 and 8 are independent of each other, as are the different sections in Chapter 8. Alternately, a one-semester course could include much of Chapters 1 to 5, Chapter 9, and the first part of Chapter 10. All topics could be covered comfortably in a two-semester course. The book can also be used to supplement courses on operator theory, operator algebras and numerical linear algebra. The book has several exercises scattered in the text and a section called Problems at the end of each chapter. An *exercise* is placed at a particular spot with the idea that the reader should do it at that stage of his reading and then proceed further. *Problems*, on the other hand, are designed to serve different purposes. Some of them are supplementary exercises, while others are about themes that are related to the main development in the text. Some are quite easy while others are hard enough to be contents of research papers. From Chapter 6 onwards, I have also used the problems for another purpose. There are results, or proofs, which are a bit too special to be placed in the main text. At the same time they are interesting enough to merit the attention of anyone working, or planning to work, in this area. I have stated such results as parts of the Problems section, often with hints about their solutions. This should enhance the value of the book as a reference, and provide topics for a seminar course as well. The reader should not be discouraged if he finds some of these problems difficult. At a few places I have drawn attention to some unsolved research problems. At some others, the existence of such problems can be inferred from the text. I hope the book will encourage some readers to solve these problems too.

While most of the notations used are the standard ones, some need a little explanation:

Almost all functional analysis books written by mathematicians adopt the convention that an inner product $\langle u, v \rangle$ is linear in the variable u and

conjugate-linear in the variable v. Physicists and numerical analysts adopt the opposite convention, and different notations as well. There would be no special reason to prefer one over the other, except that certain calculations and manipulations become much simpler in the latter notation. If u and v are column vectors, then u^*v is the product of a row vector and a column vector, hence a number. This is the inner product of u and v. Combined with the usual rules of matrix multiplication, this facilitates computations. For this reason, I have chosen the second convention about inner products, with the belief that the initial discomfort this causes some readers will be offset by the eventual advantages. (Dirac's bra and ket notation, used by physicists, is different typographically but has the same idea behind it.)

The k-fold tensor power of an operator is represented in this book as $\otimes^k A$, the antisymmetric and the symmetric tensor powers as $\wedge^k A$ and $\vee^k A$, respectively. This helps in thinking of these objects as maps, $A \to \otimes^k A$, etc. We often study the variational behaviour of, and perturbation bounds for, functions of operators. In such contexts, this notation is natural.

Very often we have to compare two n-tuples of numbers after rearranging them. For this I have used a pictorial notation that makes it easy to remember the order that has been chosen. If $x = (x_1, \ldots, x_n)$ is a vector with real coordinates, then $x^{\downarrow}$ and $x^{\uparrow}$ are vectors whose coordinates are obtained by rearranging the numbers x_j in decreasing order and in increasing order, respectively. We write $x^{\downarrow} = (x_1^{\downarrow}, \ldots, x_n^{\downarrow})$ and $x^{\uparrow} = (x_1^{\uparrow}, \ldots, x_n^{\uparrow})$, where $x_1^{\downarrow} \geq \cdots \geq x_n^{\downarrow}$ and $x_1^{\uparrow} \leq \cdots \leq x_n^{\uparrow}$.

The symbol $|||\cdot|||$ stands for a unitarily invariant norm on matrices: one that satisfies the equality $|||UAV||| = |||A|||$ for all A and for all unitary U, V. A statement like $|||A||| \leq |||B|||$ means that, for the matrices A and B, this inequality is true simultaneously for all unitarily invariant norms. The supremum norm of A, as an operator on the space $\mathbb{C}^n$, is always written as $||A||$. Other norms carry special subscripts. For example, the Frobenius norm, or the Hilbert-Schmidt norm, is written as $||A||_2$. (This should be noted by numerical analysts who often use the symbol $||A||_2$ for what we call $||A||$.)

A few symbols have different meanings in different contexts. The reader's attention is drawn to three such symbols. If x is a complex number, $|x|$ denotes the absolute value of x. If x is an n-vector with coordinates $(x_1, \ldots, x_n)$, then $|x|$ is the vector $(|x_1|, \ldots, |x_n|)$. For a matrix A, the symbol $|A|$ stands for the positive semidefinite matrix $(A^*A)^{1/2}$. If J is a finite set, $|J|$ denotes the number of elements of J. A permutation on n indices is often denoted by the symbol σ. In this case, $\sigma(j)$ is the image of the index j under the map σ. For a matrix $A, \sigma(A)$ represents the spectrum of A. The trace of a matrix A is written as tr A. In analogy, if $x = (x_1, \ldots, x_n)$ is a vector, we write tr x for the sum Σx_j.

The words matrix and operator are used interchangeably in the book. When a statement about an operator is purely finite-dimensional in content,

I use the word matrix. If a statement is true also in infinite-dimensional spaces, possibly with a small modification, I use either the word matrix or the word operator. Many of the theorems in this book have extensions to infinite-dimensional spaces.

Several colleagues have contributed to this book, directly and indirectly. I am thankful to all of them. T. Ando, J.S. Aujla, R.B. Bapat, A. Ben Israel, I. Ionascu, A.K. Lal, R.-C.Li, S.K. Narayan, D. Petz and P. Rosenthal read parts of the manuscript and brought several errors to my attention. Fumio Hiai read the whole book with his characteristic meticulous attention and helped me eliminate many mistakes and obscurities. Long-time friends and coworkers M.D. Choi, L. Elsner, J.A.R. Holbrook, R. Horn, F. Kittaneh, A. McIntosh, K. Mukherjea, K.R. Parthasarathy, P. Rosenthal and K.B. Sinha, have generously shared with me their ideas and insights. These ideas, collected over the years, have influenced my writing.

I owe a special debt to T. Ando. I first learnt some of the topics presented here from his Hokkaido University lecture notes. I have also learnt much from discussions and correspondence with him. I have taken a lot from his notes while writing this book.

The idea of writing this book came from Chandler Davis in 1986. Various logistic difficulties forced us to abandon our original plans of writing it together. The book is certainly the poorer for it. Chandler, however, has contributed so much to my mathematics, to my life, and to this project, that this is as much his book as it is mine.

I am thankful to the Indian Statistical Institute, whose facilities have made it possible to write this book. I am also thankful to the Department of Mathematics of the University of Toronto and to NSERC Canada, for several visits that helped this project take shape.

It is a pleasure to thank V.P. Sharma for his LaTeXtyping, done with competence and with good cheer, and the staff at Springer-Verlag for their help and support.

My most valuable resource while writing, has been the unstinting and ungrudging support from my son Gautam and wife Irpinder. Without that, this project might have been postponed indefinitely.

Rajendra Bhatia

Contents

I

A Review of Linear Algebra

In this chapter we review, at a brisk pace, the basic concepts of linear and multilinear algebra. Most of the material will be familiar to a reader who has had a standard Linear Algebra course, so it is presented quickly with no proofs. Some topics, like tensor products, might be less familiar. These are treated here in somewhat greater detail. A few of the topics are quite advanced and their presentation is new.

I.1 Vector Spaces and Inner Product Spaces

Throughout this book we will consider finite-dimensional vector spaces over the field $\mathbb{C}$ of complex numbers. Such spaces will be denoted by symbols V, W, V_1, V_2, etc. Vectors will, most often, be represented by symbols u, v, w, x, etc., and scalars by a, b, s, t, etc. The symbol n, when not explained, will always mean the dimension of the vector space under consideration.

Most often, our vector space will be an inner product space. The inner product between the vectors u, v will be denoted by $\langle u, v\rangle$. We will adopt the convention that this is conjugate-linear in the first variable u and linear in the second variable v. We will always assume that the inner product is definite; i.e., $\langle u, u\rangle = 0$ if and only if $u = 0$. A vector space with such an inner product is then a finite-dimensional Hilbert space. Spaces of this type will be denoted by symbols $\mathcal{H}, \mathcal{K}$, etc. The norm arising from the inner product will be denoted by $\|u\|$; i.e., $\|u\| = \langle u, u\rangle^{1/2}$.

As usual, it will sometimes be convenient to deal with the standard Hilbert space $\mathbb{C}^n$. Elements of this vector space are column vectors with

n coordinates. In this case, the inner product $\langle u, v\rangle$ is the matrix product u^*v obtained by multiplying the column vector v on the left by the row vector u^*. The symbol $*$ denotes the conjugate transpose for matrices of any size. The notation u^*v for the inner product is sometimes convenient even when the Hilbert space is not $\mathbb{C}^n$.

The distinction between column vectors and row vectors is important in manipulations involving products. For example, if we write elements of $\mathbb{C}^n$ as column vectors, then u^*v is a number, but uv^* is an $n \times n$ matrix (sometimes called the "outer product" of u and v). However, it is typographically inconvenient to write column vectors. So, when the context does not demand this distinction, we may write a vector x with scalar coordinates $x_1, \ldots, x_n$, simply as $(x_1, \ldots, x_n)$. This will often be done in later chapters. For the present, however, we will maintain the distinction between row and column vectors.

Occasionally our Hilbert spaces will be real, but we will use the same notation for them as for the complex ones. Many of our results will be true for infinite-dimensional Hilbert spaces, with appropriate modifications at times. We will mention this only in passing.

Let $X = (x_1, \ldots, x_k)$ be a k-tuple of vectors. If these are column vectors, then X is an $n \times k$ matrix. This notation suggests matrix manipulations with X that are helpful even in the general case.

For example, let $X = (x_1, \ldots, x_k)$ be a linearly independent k-tuple. We say that a k-tuple $Y = (y_1, \ldots, y_k)$ is **biorthogonal** to X if $\langle y_i, x_j\rangle = \delta_{ij}$. This condition is expressed in matrix terms as $Y^*X = I_k$, the $k \times k$ identity matrix.

Exercise I.1.1 *Given any k-tuple of linearly independent vectors X as above, there exists a k-tuple Y biorthogonal to it. If $k = n$, this Y is unique.*

The Gram-Schmidt procedure, in this notation, can be interpreted as a matrix factoring theorem. Given an n-tuple $X = (x_1, \ldots, x_n)$ of linearly independent vectors the procedure gives another n-tuple $Q = (q_1, \ldots, q_n)$ whose entries are orthonormal vectors. For each $k = 1, 2, \ldots, n$, the vectors $\{x_1, \ldots, x_k\}$ and $\{q_1, \ldots, q_k\}$ have the same linear span. In matrix notation this can be expressed as an equation, $X = QR$, where R is an upper triangular matrix. The matrix R may be chosen so that all its diagonal entries are positive. With this restriction the factors Q and R are both unique. If the vectors x_j are not linearly independent, this procedure can be modified. If the vector x_k is linearly dependent on $x_1, \ldots, x_{k-1}$, set $q_k = 0$; otherwise proceed as in the Gram-Schmidt process. If the kth column of the matrix Q so constructed is zero, put the kth row of R to be zero. Now we have a factorisation $X = QR$, where R is upper triangular and Q has orthogonal columns, some of which are zero. Take the nonzero columns of Q and extend this set to an orthonormal basis. Then, replace the zero columns of Q by these additional basis vectors. The new matrix Q now has orthonormal columns, and we still have $X = QR$, because the

new columns of Q are matched with zero rows of R. This is called the **QR decomposition**.

Similarly, a change of orthogonal bases can be conveniently expressed in these notations as follows. Let $X = (x_1, \ldots, x_k)$ be any k-tuple of vectors and $E = (e_1, \ldots, e_n)$ any orthonormal basis. Then, the columns of the matrix E^*X are the representations of the vectors comprising X, relative to the basis E. When $k = n$ and X is an orthonormal basis, then E^*X is a unitary matrix. Furthermore, this is the matrix by which we pass between coordinates of any vector relative to the basis E and those relative to the basis X. Indeed, if

$$u = a_1e_1 + \cdots + a_ne_n = b_1x_1 + \cdots + b_nx_n,$$

then we have

$$\begin{array}{lll} u = Ea, & a_j = e_j^*u, & a = E^*u, \\ u = Xb, & b_j = x_j^*u, & b = X^*u. \end{array}$$

Hence,

$$a = E^*Xb \quad \text{and} \quad b = X^*Ea.$$

Exercise I.1.2 *Let X be any basis of $\mathcal{H}$ and let Y be the basis biorthogonal to it. Using matrix multiplication, X gives a linear transformation from $\mathbb{C}^n$ to $\mathcal{H}$. The inverse of this is given by Y^*. In the special case when X is orthonormal (so that $Y = X$), this transformation is inner-product-preserving if the standard inner product is used on $\mathbb{C}^n$.*

Exercise I.1.3 *Use the QR decomposition to prove* **Hadamard's inequality**: *if $X = (x_1, \ldots, x_n)$, then*

$$|\det X| \leq \prod_{j=1}^{n} \|x_j\|.$$

Equality holds here if and only if either the x_j are mutually orthogonal or some x_j is zero.

I.2 Linear Operators and Matrices

Let $\mathcal{L}(V, W)$ be the space of all linear operators from a vector space V to a vector space W. If bases for V, W are fixed, each such operator has a unique matrix associated with it. As usual, we will talk of operators and matrices interchangeably.

For operators between Hilbert spaces, the matrix representations are especially nice if the bases chosen are orthonormal. Let $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, and let $E = (e_1, \ldots, e_n)$ be an orthonormal basis of $\mathcal{H}$ and $F = (f_1, \ldots, f_m)$ an orthonormal basis of $\mathcal{K}$. Then, the (i, j)-entry of the matrix of A relative

to these bases is $a_{ij} = f_i^* A e_j = \langle f_i, Ae_j \rangle$. This suggests that we may say that the matrix of A relative to these bases is $F^* A E$.

In this notation, composition of linear operators can be identified with matrix multiplication as follows. Let $\mathcal{M}$ be a third Hilbert space with orthonormal basis $G = (g_1, \ldots, g_p)$. Let $B \in \mathcal{L}(\mathcal{K}, \mathcal{M})$. Then

$$\begin{aligned}
(\text{matrix of } B \cdot A) &= G^*(B \cdot A)E \\
&= G^* B F \, F^* A E \\
&= (G^* B F)(F^* A E) \\
&= (\text{matrix of } B)\ (\text{matrix of } A).
\end{aligned}$$

The second step in the above chain is justified by Exercise I.1.2.

The **adjoint** of an operator $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ is the unique operator A^* in $\mathcal{L}(\mathcal{K}, \mathcal{H})$ that satisfies the relation

$$\langle z, Ax \rangle_{\mathcal{K}} = \langle A^* z, x \rangle_{\mathcal{H}}$$

for all $x \in \mathcal{H}$ and $z \in \mathcal{K}$.

Exercise I.2.1 *For fixed bases in $\mathcal{H}$ and $\mathcal{K}$, the matrix of A^* is the conjugate transpose of the matrix of A.*

For the space $\mathcal{L}(\mathcal{H}, \mathcal{H})$ we use the more compact notation $\mathcal{L}(\mathcal{H})$. In the rest of this section, and elsewhere in the book, if no qualification is made, an operator would mean an element of $\mathcal{L}(\mathcal{H})$.

An operator A is called **self-adjoint** or **Hermitian** if $A = A^*$, **skew-Hermitian** if $A = -A^*$, **unitary** if $AA^* = I = A^*A$, and **normal** if $AA^* = A^*A$.

A Hermitian operator A is said to be **positive** or **positive semidefinite** if $\langle x, Ax \rangle \geq 0$ for all $x \in \mathcal{H}$. The notation $A \geq 0$ will be used to express the fact that A is a positive operator. If $\langle x, Ax \rangle > 0$ for all nonzero x, we will say A is **positive definite**, or **strictly positive** . We will then write $A > 0$. A positive operator is strictly positive if and only if it is invertible. If A and B are Hermitian, then we say $A \geq B$ if $A - B \geq 0$.

Given any operator A we can find an orthonormal basis $y_1, \ldots, y_n$ such that for each $k = 1, 2, \ldots, n$, the vector Ay_k is a linear combination of $y_1, \ldots, y_k$. This can be proved by induction on the dimension n of $\mathcal{H}$. Let λ_1 be any eigenvalue of A and y_1 an eigenvector corresponding to λ_1, and $\mathcal{M}$ the 1-dimensional subspace spanned by it. Let $\mathcal{N}$ be the orthogonal complement of $\mathcal{M}$. Let $P_{\mathcal{N}}$ denote the orthogonal projection on $\mathcal{N}$. For $y \in \mathcal{N}$, let $A_{\mathcal{N}} y = P_{\mathcal{N}} A y$. Then, $A_{\mathcal{N}}$ is a linear operator on the $(n-1)$-dimensional space $\mathcal{N}$. So, by the induction hypothesis, there exists an orthogonal basis $y_2, \ldots, y_n$ of $\mathcal{N}$ such that for $k = 2, \ldots, n$ the vector $A_{\mathcal{N}} y_k$ is a linear combination of $y_2, \ldots, y_k$. Now $y_1, \ldots, y_n$ is an orthogonal basis for $\mathcal{H}$, and each Ay_k is a linear combination of $y_1, \ldots, y_k$ for $k = 1, 2, \ldots, n$. Thus, the matrix of A with respect to this basis is upper triangular. In other words,

every matrix A is **unitarily equivalent** (or **unitarily similar**) to an upper triangular matrix T, i.e., $A = QTQ^*$, where Q is unitary and T is upper triangular. This triangular matrix is called a **Schur Triangular Form** for A. An orthonormal basis with respect to which A is upper triangular is called a **Schur basis** for A. If A is normal, then T is diagonal and we have $Q^*AQ = D$, where D is a diagonal matrix whose diagonal entries are the eigenvalues of A. This is the **Spectral Theorem** for normal matrices.

The Spectral Theorem makes it easy to define functions of normal matrices. If f is any complex function, and if D is a diagonal matrix with $\lambda_1, \ldots,$ λ_n on its diagonal, then $f(D)$ is the diagonal matrix with $f(\lambda_1), \ldots, f(\lambda_n)$ on its diagonal. If $A = QDQ^*$, then $f(A) = Qf(D)Q^*$. A special consequence, used very often, is the fact that every positive operator A has a unique positive square root. This square root will be written as $A^{1/2}$.

Exercise I.2.2 *Show that the following statements are equivalent:*

(i) A is positive.

(ii) $A = B^*B$ for some B.

(iii) $A = T^*T$ for some upper triangular T.

(iv) $A = T^*T$ for some upper triangular T with nonnegative diagonal entries.

If A is positive definite, then the factorisation in (iv) is unique. This is called the **Cholesky Decomposition** *of A.*

Exercise I.2.3 *(i) Let $\{A_\alpha\}$ be a family of mutually commuting operators. Then, there is a common Schur basis for $\{A_\alpha\}$. In other words, there exists a unitary Q such that $Q^*A_\alpha Q$ is upper triangular for all α.*

(ii) Let $\{A_\alpha\}$ be a family of mutually commuting normal operators. Then, there exists a unitary Q such that $Q^*A_\alpha Q$ is diagonal for all α.

For any operator A the operator A^*A is always positive, and its unique positive square root is denoted by $|A|$. The eigenvalues of $|A|$ counted with multiplicities are called the **singular values** of A. We will always enumerate these in decreasing order, and use for them the notation $s_1(A) \geq s_2(A) \geq \cdots \geq s_n(A)$.

If rank $A = k$, then $s_k(A) > 0$, but $s_{k+1}(A) = \cdots = s_n(A) = 0$. Let S be the diagonal matrix with diagonal entries $s_1(A), \ldots, s_n(A)$ and S_+ the $k \times k$ diagonal matrix with diagonal entries $s_1(A), \ldots, s_k(A)$. Let $Q = (Q_1, Q_2)$ be the unitary matrix in which Q_1 is the $n \times k$ matrix whose columns are the eigenvectors of A^*A corresponding to the eigenvalues $s_1^2(A), \ldots, s_k^2(A)$ and Q_2 the $n \times (n-k)$ matrix whose columns are the eigenvectors of A^*A corresponding to the remaining eigenvalues. Then, by the Spectral Theorem

$$Q^*(A^*A)Q = \begin{pmatrix} S_+^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that

$$Q_1^*(A^*A)Q_1 = S_+^2, \quad Q_2^*(A^*A)Q_2 = 0.$$

The second of these relations implies that $AQ_2 = 0$. From the first one we can conclude that if $W_1 = AQ_1S_+^{-1}$, then $W_1^*W_1 = I_k$. Choose W_2 so that $W = (W_1, W_2)$ is unitary. Then, we have

$$W^*AQ = \begin{pmatrix} W_1^*AQ_1 & W_1^*AQ_2 \\ W_2^*AQ_1 & W_2^*AQ_2 \end{pmatrix} = \begin{pmatrix} S_+ & 0 \\ 0 & 0 \end{pmatrix}.$$

This is the **Singular Value Decomposition:** for every matrix A there exist unitaries W and Q such that

$$W^*AQ = S,$$

where S is the diagonal matrix whose diagonal entries are the singular values of A.

Note that in the above representation the columns of Q are eigenvectors of A^*A and the columns of W are eigenvectors of AA^* corresponding to the eigenvalues $s_j^2(A)$, $1 \leq j \leq n$. These eigenvectors are called the **right** and **left singular vectors** of A, respectively.

Exercise I.2.4 *(i) The Singular Value Decomposition leads to the* **Polar Decomposition:** *Every operator A can be written as $A = UP$, where U is unitary and P is positive. In this decomposition the positive part P is unique, $P = |A|$. The unitary part U is unique if A is invertible.*

(ii) An operator A is normal if and only if the factors U and P in the polar decomposition of A commute.

(iii) We have derived the Polar Decomposition from the Singular Value Decomposition. Show that it is possible to derive the latter from the former.

Every operator A can be decomposed as a sum

$$A = \operatorname{Re} A + i \operatorname{Im} A,$$

where $\operatorname{Re} A = \frac{A+A^*}{2}$ and $\operatorname{Im} A = \frac{A-A^*}{2i}$. This is called the **Cartesian Decomposition** of A into its "real" and "imaginary" parts. The operators $\operatorname{Re} A$ and $\operatorname{Im} A$ are both Hermitian.

The **norm** of an operator A is defined as

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

We also have

$$\|A\| = \sup_{\|x\|=\|y\|=1} |\langle y, Ax\rangle|.$$

When A is Hermitian we have

$$\|A\| = \sup_{\|x\|=1} |\langle x, Ax\rangle|.$$

For every operator A we have

$$\|A\| = s_1(A) = \|A^*A\|^{1/2}.$$

When A is normal we have

$$\|A\| = \max\{|\lambda_j| : \lambda_j \text{ is an eigenvalue of } A\}.$$

An operator A is said to be a **contraction** if $\|A\| \leq 1$. We also use the adjective **contractive** for such an operator. A positive operator A is contractive if and only if $A \leq I$.

To distinguish it from other norms that we consider later, the norm $\|A\|$ will be called the **operator norm** or the **bound norm** of A.

Another useful norm is the norm

$$\|A\|_2 = (\sum_{j=1}^{n} s_j^2(A))^{1/2} = (\mathrm{tr} A^*A)^{1/2},$$

where tr stands for the trace of an operator. If a_{ij} are the entries of a matrix representation of A relative to an orthonormal basis of $\mathcal{H}$, then

$$\|A\|_2 = (\sum_{i,j} |a_{ij}|^2)^{1/2}.$$

This makes this norm useful in calculations with matrices. This is called the **Frobenius norm** or the **Schatten 2-norm** or the **Hilbert-Schmidt norm**.

Both $\|A\|$ and $\|A\|_2$ have an important invariance property called **unitary invariance**: we have $\|A\| = \|UAV\|$ and $\|A\|_2 = \|UAV\|_2$ for all unitary U, V.

Any two norms on a finite-dimensional space are equivalent. For the norms $\|A\|$ and $\|A\|_2$ it follows from the properties listed above that

$$\|A\| \leq \|A\|_2 \leq n^{1/2}\|A\|$$

for every A.

Exercise I.2.5 *Show that matrices with distinct eigenvalues are dense in the space of all $n \times n$ matrices. (Use the Schur Triangularisation.)*

Exercise I.2.6 *If $\|A\| < 1$, then $I - A$ is invertible and*

$$(I - A)^{-1} = I + A + A^2 + \cdots,$$

a convergent power series. This is called the **Neumann Series**.

Exercise I.2.7 *The set of all invertible matrices is a dense open subset of the set of all $n \times n$ matrices. The set of all unitary matrices is a compact subset of the set of all $n \times n$ matrices. These two sets are also groups under multiplication. They are called the* **general linear group GL(n)** *and the* **unitary group U(n)**, *respectively.*

Exercise I.2.8 *For any matrix A the series*

$$\exp A = I + A + \frac{A^2}{2!} + \cdots + \frac{A^n}{n!} + \cdots$$

converges. This is called the **exponential** *of A. The matrix* $\exp A$ *is always invertible and*

$$(\exp A)^{-1} = \exp(-A).$$

Conversely, every invertible matrix can be expressed as the exponential of some matrix. Every unitary matrix can be expressed as the exponential of a skew-Hermitian matrix.

The **numerical range** or the **field of values** of an operator A is the subset $W(A)$ of the complex plane defined as

$$W(A) = \{\langle x, Ax\rangle : \|x\| = 1\}.$$

Note that

$$\begin{array}{rcll} W(UAU^*) & = & W(A) & \text{for all } U \in U(n), \\ W(aA + bI) & = & aW(A) + bW(I) & \text{for all } a, b \in \mathbb{C}. \end{array}$$

It is clear that if λ is an eigenvalue of A, then λ is in $W(A)$. It is also clear that $W(A)$ is a closed set. An important property of $W(A)$ is that it is a convex set. This is called the **Toeplitz-Hausdorff Theorem**; an outline of its proof is given in Problem I.6.2.

Exercise I.2.9 *(i) When A is normal, the set W(A) is the convex hull of the eigenvalues of A. For nonnormal matrices, W(A) may be bigger than the convex hull of its eigenvalues. For Hermitian operators, the first statement says that W(A) is the closed interval whose endpoints are the smallest and the largest eigenvalues of A.*

(ii) If a unit vector x belongs to the linear span of the eigenspaces corresponding to eigenvalues $\lambda_1, \ldots, \lambda_k$ *of a normal operator A, then* $\langle x, Ax\rangle$ *lies in the convex hull of* $\lambda_1, \ldots, \lambda_k$. *(This fact will be used frequently in Chapter III.)*

The number $w(A)$ defined as

$$w(A) = \sup_{\|x\|=1} |\langle x, Ax\rangle|$$

is called the **numerical radius** of A.

Exercise I.2.10 *(i) The numerical radius defines a norm on* $\mathcal{L}(\mathcal{H})$.

(ii) $w(UAU^*) = w(A)$ *for all* $U \in U(n)$.

(iii) $w(A) \leq \|A\| \leq 2w(A)$ *for all A.*

(iv) $w(A) = \|A\|$ if (but not only if) A is normal.

The **spectral radius** of an operator A is defined as

$$\operatorname{spr}(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

We have noted that $\operatorname{spr}(A) \le w(A) \le \|A\|$, and that the three are equal if (but not only if) the operator A is normal.

I.3 Direct Sums

If U, V are vector spaces, their **direct sum** is the space of columns $\binom{u}{v}$ with $u \in U$ and $v \in V$. This is a vector space with vector operations naturally defined coordinatewise. If $\mathcal{H}, \mathcal{K}$ are Hilbert spaces, their direct sum is a Hilbert space with inner product defined as

$$\left\langle \begin{pmatrix} h \\ k \end{pmatrix}, \begin{pmatrix} h' \\ k' \end{pmatrix} \right\rangle = \langle h, h' \rangle_{\mathcal{H}} + \langle k, k' \rangle_{\mathcal{K}}.$$

We will always denote this direct sum as $\mathcal{H} \oplus \mathcal{K}$.

If $\mathcal{M}$ and $\mathcal{N}$ are orthogonally complementary subspaces of $\mathcal{H}$, then the fact that every vector x in $\mathcal{H}$ has a unique representation $x = u + v$ with $u \in \mathcal{M}$ and $v \in \mathcal{N}$ implies that $\mathcal{H}$ is isomorphic to $\mathcal{M} \oplus \mathcal{N}$. This isomorphism is given by a natural, fixed map. So, we say that $\mathcal{H} = \mathcal{M} \oplus \mathcal{N}$. When a distinction is necessary we call this an **internal direct sum**. If $\mathcal{M}, \mathcal{N}$ are subspaces of $\mathcal{H}$ complementary in the algebraic but not in the orthogonal sense; i.e., if $\mathcal{M}$ and $\mathcal{N}$ are disjoint and their linear span is $\mathcal{H}$, then every vector x in $\mathcal{H}$ has a unique decomposition $x = u + v$ as before, but not with orthogonal summands. In this case we write $\mathcal{H} \doteq \mathcal{M} + \mathcal{N}$ and say $\mathcal{H}$ is the **algebraic direct sum** of $\mathcal{M}$ and $\mathcal{N}$.

If $\mathcal{H} = \mathcal{M} \oplus \mathcal{N}$ is an internal direct sum, we may define the injection of $\mathcal{M}$ into $\mathcal{H}$ as the operator $I_{\mathcal{M}} \in \mathcal{L}(\mathcal{M}, \mathcal{H})$ such that $I_{\mathcal{M}}(u) = u$ for all $u \in \mathcal{M}$. Then, $I_{\mathcal{M}}^*$ is an element of $\mathcal{L}(\mathcal{H}, \mathcal{M})$ defined as $I_{\mathcal{M}}^* x = Px$ for all $x \in \mathcal{H}$, where P is the orthoprojector onto $\mathcal{M}$. Here one should note that $I_{\mathcal{M}}^*$ is not the same as P because they map into different spaces. That is why their adjoints can be different. Similarly define $I_{\mathcal{N}}$. Then, $(I_{\mathcal{M}}, I_{\mathcal{N}})$ is an isometry from the ordinary ("external") direct sum $\mathcal{M} \oplus \mathcal{N}$ onto $\mathcal{H}$.

If $\mathcal{H} = \mathcal{M} \oplus \mathcal{N}$ and $A \in \mathcal{L}(\mathcal{H})$, then using this isomorphism, we can write A as a **block-matrix**

$$A = \begin{pmatrix} B & C \\ D & E \end{pmatrix},$$

where $B \in \mathcal{L}(\mathcal{M}), C \in \mathcal{L}(\mathcal{N}, \mathcal{M})$, etc. Here, for example, $C = I_{\mathcal{M}}^* A I_{\mathcal{N}}$. The usual rules of matrix operations hold for block matrices. Adjoints are obtained by taking "conjugate transposes" formally.

If the subspace $\mathcal{M}$ is invariant under A; i.e., $Ax \in \mathcal{M}$ whenever $x \in \mathcal{M}$, then in the above block-matrix representation of A we must have $D = 0$. Indeed, this condition is equivalent to $\mathcal{M}$ being invariant. If both $\mathcal{M}$ and its orthogonal complement $\mathcal{N}$ are invariant under A, we say that $\mathcal{M}$ **reduces** A. In this case, both C and D are 0. We then say that the operator A is the direct sum of B and E and write $A = B \oplus E$.

Exercise I.3.1 *Let $A = A_1 \oplus A_2$. Show that*

(i) $W(A)$ is the convex hull of $W(A_1)$ and $W(A_2)$; i.e., the smallest convex set containing $W(A_1) \cup W(A_2)$.

(ii)
$$\begin{aligned} \|A\| &= \max(\|A_1\|, \|A_2\|), \\ \operatorname{spr}(A) &= \max(\operatorname{spr}(A_1), \operatorname{spr}(A_2)), \\ w(A) &= \max(w(A_1), w(A_2)). \end{aligned}$$

Direct sums in which each summand $\mathcal{H}_j$ is the same space $\mathcal{H}$ arise often in practice. Very often, some properties of an operator A on $\mathcal{H}$ are reflected in those of some other operators on $\mathcal{H} \oplus \mathcal{H}$. This is illustrated in the following propositions.

Lemma I.3.2 *Let $A \in \mathcal{L}(\mathcal{H})$. Then, the operators $\begin{pmatrix} A & A \\ A & A \end{pmatrix}$ and $\begin{pmatrix} 2A & 0 \\ 0 & 0 \end{pmatrix}$ are unitarily equivalent in $\mathcal{L}(\mathcal{H} \oplus \mathcal{H})$.*

Proof. The equivalence is implemented by the unitary operator $\frac{1}{\sqrt{2}}\begin{pmatrix} I & I \\ -I & I \end{pmatrix}$. ■

Corollary I.3.3 *An operator A on $\mathcal{H}$ is positive if and only if the operator $\begin{pmatrix} A & A \\ A & A \end{pmatrix}$ on $\mathcal{H} \oplus \mathcal{H}$ is positive.*

This can also be seen by writing $\begin{pmatrix} A & A \\ A & A \end{pmatrix} = \begin{pmatrix} A^{1/2} & 0 \\ A^{1/2} & 0 \end{pmatrix} \begin{pmatrix} A^{1/2} & A^{1/2} \\ 0 & 0 \end{pmatrix}$, and using Exercise I.2.2.

Corollary I.3.4 *For every $A \in \mathcal{L}(\mathcal{H})$ the operator $\begin{pmatrix} |A| & A^* \\ A & |A^*| \end{pmatrix}$ is positive.*

Proof. Let $A = UP$ be the polar decomposition of A. Then,

$$\begin{aligned} \begin{pmatrix} |A| & A^* \\ A & |A^*| \end{pmatrix} &= \begin{pmatrix} P & PU^* \\ UP & UPU^* \end{pmatrix} \\ &= \begin{pmatrix} I & O \\ O & U \end{pmatrix} \begin{pmatrix} P & P \\ P & P \end{pmatrix} \begin{pmatrix} I & O \\ O & U^* \end{pmatrix}. \end{aligned}$$

Note that $\begin{pmatrix} I & O \\ O & U \end{pmatrix}$ is a unitary operator on $\mathcal{H} \oplus \mathcal{H}$. ■

Proposition I.3.5 *An operator A on $\mathcal{H}$ is contractive if and only if the operator $\begin{pmatrix} I & A^* \\ A & I \end{pmatrix}$ on $\mathcal{H} \oplus \mathcal{H}$ is positive.*

Proof. If A has the singular value decomposition $A = USV^*$, then

$$\begin{pmatrix} I & A^* \\ A & I \end{pmatrix} = \begin{pmatrix} V & O \\ O & U \end{pmatrix} \begin{pmatrix} I & S \\ S & I \end{pmatrix} \begin{pmatrix} V^* & O \\ O & U^* \end{pmatrix}.$$

Hence $\begin{pmatrix} I & A^* \\ A & I \end{pmatrix}$ is positive if and only if $\begin{pmatrix} I & S \\ S & I \end{pmatrix}$ is positive. Also, $\|A\| = \|S\|$. So we may assume, without loss of generality, that $A = S$.

Now let W be the unitary operator on $\mathcal{H} \oplus \mathcal{H}$ that sends the orthonormal basis $\{e_1, e_2, \ldots, e_{2n}\}$ to the basis $\{e_1, e_{n+1}, e_2, e_{n+2}, \ldots, e_n, e_{2n}\}$. Then, the unitary conjugation by W transforms the matrix $\begin{pmatrix} I & S \\ S & I \end{pmatrix}$ to a direct sum of n two-by-two matrices

$$\begin{pmatrix} 1 & s_1 \\ s_1 & 1 \end{pmatrix} \oplus \begin{pmatrix} 1 & s_2 \\ s_2 & 1 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 1 & s_n \\ s_n & 1 \end{pmatrix}.$$

This is positive if and only if each of the summands is positive, which happens if and only if $s_j \leq 1$ for all j; i.e., S is a contraction. ■

Exercise I.3.6 *If A is a contraction, show that*

$$A^*(I - AA^*)^{1/2} = (I - A^*A)^{1/2}A^*.$$

Use this to show that if A is a contraction on $\mathcal{H}$, then the operators

$$U = \begin{pmatrix} A & (I - AA^*)^{1/2} \\ (I - A^*A)^{1/2} & -A^* \end{pmatrix},$$
$$V = \begin{pmatrix} A & -(I - AA^*)^{1/2} \\ (I - A^*A)^{1/2} & A^* \end{pmatrix}$$

are unitary operators on $\mathcal{H} \oplus \mathcal{H}$.

Exercise I.3.7 *For every matrix A, the matrix $\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}$ is invertible and its inverse is $\begin{pmatrix} I & -A \\ 0 & I \end{pmatrix}$. Use this to show that if A, B are any two $n \times n$ matrices, then*

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}.$$

This implies that AB and BA have the same eigenvalues. (This last fact can be proved in another way as follows. If B is invertible, then $AB = B^{-1}(BA)B$. So, AB and BA have the same eigenvalues. Since invertible matrices are dense in the space of all matrices, and a general known fact in complex analysis is that the roots of a polynomial vary continuously with the coefficients, the above conclusion also holds in general.)

Direct sums with more than two summands are defined in the same way. We will denote the direct sum of spaces $\mathcal{H}_1, \ldots, \mathcal{H}_k$ as $\oplus_{j=1}^k \mathcal{H}_j$, or simply as $\oplus_j \mathcal{H}_j$.

I.4 Tensor Products

Let $V_j, 1 \leq j \leq k$, be vector spaces. A map F from the Cartesian product $V_1 \times \cdots \times V_k$ to another vector space W is called **multilinear** if it depends linearly on each of the arguments. When $W = \mathbb{C}$, such maps are called **multilinear functionals**. When $k = 2$, the word multilinear is replaced by **bilinear**. Bilinear maps, thus, are maps $F : V_1 \times V_2 \to W$ that satisfy the conditions

$$\begin{aligned} F(u, av_1 + bv_2) &= aF(u, v_1) + bF(u, v_2), \\ F(au_1 + bu_2, v) &= aF(u_1, v) + bF(u_2, v), \end{aligned}$$

for all $a, b \in \mathbb{C}$; $u, u_1, u_2 \in V_1$ and $v, v_1, v_2 \in V_2$. We will be looking most often at the special situation when each V_j is the same vector space.

As a special example consider a Hilbert space $\mathcal{H}$ and fix two vectors x, y in it. Then,

$$F(u, v) = \langle x, u \rangle \langle y, v \rangle$$

is a bilinear functional on $\mathcal{H}$.

We see from this example that it is equally natural to consider conjugate-multilinear functionals as well. Even more generally we could study functions that are linear in some variables and conjugate-linear in others. As an example, let $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ and for $u \in \mathcal{K}$ and $v \in \mathcal{H}$, let $F(u, v) = \langle u, Av \rangle_{\mathcal{K}}$. Then, F depends linearly on v and conjugate-linearly on u. Such functionals are called **sesquilinear**; an inner product is a functional of this sort. The example given above is the "most general" example of a sesquilinear functional: if $F(u, v)$ is any sesquilinear functional on $\mathcal{K} \times \mathcal{H}$, then there exists a unique operator $A \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ such that $F(u, v) = \langle u, Av \rangle$.

In this sense our first example is not the most general example of a bilinear functional. Bilinear functionals $F(u, v)$ on $\mathcal{H}$ that can be expressed as $F(u, v) = \langle x, u \rangle \langle y, v \rangle$ for some fixed $x, y \in \mathcal{H}$ are called **elementary**. They are special as the following exercise will show.

Exercise I.4.1 *Let x, y, z be linearly independent vectors in $\mathcal{H}$. Find a necessary and sufficient condition that a vector w must satisfy in order that the bilinear functional*

$$F(u, v) = \langle x, u \rangle \langle y, v \rangle + \langle z, u \rangle \langle w, v \rangle$$

is elementary.

The set of all bilinear functionals is a vector space. The result of this exercise shows that the subset consisting of elementary functionals is not closed under addition. We will soon see that a convenient basis for this vector space can be constructed with elementary functionals as its members.

The procedure, called the tensor product construction, starts by taking formal linear combinations of symbols $x \otimes y$ with $x \in \mathcal{H}, y \in \mathcal{K}$; then

reducing this space modulo suitable equivalence relations; then identifying the resulting space with the space of bilinear functionals.

More precisely, consider all finite sums of the type $\sum_i c_i(x_i \otimes y_i)$, $c_i \in \mathbb{C}, x_i \in \mathcal{H}, y_i \in \mathcal{K}$ and manipulate them formally as linear combinations. In this space the expressions

$$\begin{array}{rcl} a(x \otimes y) & - & (ax \otimes y) \\ a(x \otimes y) & - & (x \otimes ay) \\ x_1 \otimes y + x_2 \otimes y & - & (x_1 + x_2) \otimes y \\ x \otimes y_1 + x \otimes y_2 & - & x \otimes (y_1 + y_2) \end{array}$$

are next defined to be equivalent to 0, for all $a \in \mathbb{C}; x, x_1, x_2 \in \mathcal{H}$ and $y, y_1, y_2 \in \mathcal{K}$. The set of all linear combinations of expressions $x \otimes y$ for $x \in \mathcal{H}, y \in \mathcal{K}$, after reduction modulo these equivalences, is called the **tensor product** of $\mathcal{H}$ and $\mathcal{K}$ and is denoted as $\mathcal{H} \otimes \mathcal{K}$.

Each term $c(x \otimes y)$ determines a conjugate-bilinear functional $F^*(u, v)$ on $\mathcal{H} \times \mathcal{K}$ by the natural rule

$$F^*(u, v) = c\langle u, x\rangle\langle v, y\rangle.$$

This can be extended to sums of such terms, and the equivalences were chosen in such a way that equivalent expressions (i.e., expressions giving the same element of $\mathcal{H}\otimes\mathcal{K}$) give the same functional. The complex conjugate of each such functional gives a bilinear functional. These ideas can be extended directly to k-linear functionals, including those that are linear in some of the arguments and conjugate-linear in others.

Theorem I.4.2 *The space of all bilinear functionals on $\mathcal{H}$ is linearly spanned by the elementary ones. If $(e_1, \ldots, e_n)$ is a fixed orthonormal basis of $\mathcal{H}$, then to every bilinear functional F there correspond unique vectors $x_1, \ldots, x_n$ such that*

$$F^* = \sum_j e_j \otimes x_j.$$

Every sequence $x_j, 1 \leq j \leq n$, leads to a bilinear functional in this way.

Proof. Let F be a bilinear functional on $\mathcal{H}$. For each $j, F^*(e_j, v)$ is a conjugate-linear function of v. Hence there exists a unique vector x_j such that $F^*(e_j, v) = \langle v, x_j\rangle$ for all v.

Now, if $u = \Sigma a_j e_j$ is any vector in $\mathcal{H}$, then $F(u, v) = \Sigma a_j F(e_j, v) = \Sigma\langle e_j, u\rangle\langle x_j, v\rangle$. In other words, $F^* = \Sigma e_j \otimes x_j$ as asserted. ■

A more symmetric form of the above statement is the following:

Corollary I.4.3 *If $(e_1, \ldots, e_n)$ and $(f_1, \ldots, f_n)$ are two fixed orthonormal bases of $\mathcal{H}$, then every bilinear functional F on $\mathcal{H}$ has a unique representation $F = \Sigma a_{ij}(e_i \otimes f_j)^*$.*

(Most often, the choice $(e_1, \ldots, e_n) = (f_1, \ldots, f_n)$ is the convenient one for using the above representations.)

Thus, it is natural to denote the space of conjugate-bilinear functionals on $\mathcal{H}$ by $\mathcal{H} \otimes \mathcal{H}$. This is an n^2-dimensional vector space. The inner product on this space is defined by putting

$$\langle u_1 \otimes u_2, v_1 \otimes v_2 \rangle = \langle u_1, v_1 \rangle \langle u_2, v_2 \rangle,$$

and then extending this definition to all of $\mathcal{H} \otimes \mathcal{H}$ in a natural way. It is easy to verify that this definition is consistent with the equivalences used in defining the tensor product. If $(e_1, \ldots, e_n)$ and $(f_1, \ldots, f_n)$ are orthonormal bases in $\mathcal{H}$, then $e_i \otimes f_j$, $1 \leq i, j \leq n$, form an orthonormal basis in $\mathcal{H} \otimes \mathcal{H}$. For the purposes of computation it is useful to order this basis **lexicographically**: we say that $e_i \otimes f_j$ precedes $e_k \otimes f_\ell$ if and only if either $i < k$ or $i = k$ and $j < \ell$.

Tensor products such as $\mathcal{H} \otimes \mathcal{K}$ or $\mathcal{K}^* \otimes \mathcal{H}$ can be defined by imitating the above procedure. Here the space $\mathcal{K}^*$ is the space of all conjugate-linear functionals on $\mathcal{K}$. This space is called the **dual space** of $\mathcal{K}$. There is a natural identification between $\mathcal{K}$ and $\mathcal{K}^*$ via a conjugate-linear, norm preserving bijection.

Exercise I.4.4 *(i) There is a natural isomorphism between the spaces $\mathcal{K} \otimes \mathcal{H}^*$ and $\mathcal{L}(\mathcal{H}, \mathcal{K})$ in which the elementary tensor $k \otimes h^*$ corresponds to the linear map that takes a vector u of $\mathcal{H}$ to $\langle h, u \rangle k$. This linear transformation has rank one and all rank one, transformations can be obtained in this way.*

(ii) An explicit construction of this isomorphism φ is outlined below. Let $e_1, \ldots, e_n$ be an orthonormal basis for $\mathcal{H}$ and for $\mathcal{H}^*$. Let $f_1, \ldots, f_m$ be an orthonormal basis for $\mathcal{K}$. Identitfy each element of $\mathcal{L}(\mathcal{H}, \mathcal{K})$ with its matrix with respect to these bases. Let E_{ij} be the matrix all whose entries are zero except the (i, j)-entry, which is 1. Show that $\varphi(f_i \otimes e_j) = E_{ij}$ for all $1 \leq i \leq m$, $1 \leq j \leq n$. Thus, if A is any $m \times n$ matrix with entries a_{ij}, then

$$\varphi^{-1}(A) = \sum_{i,j} a_{ij}(f_i \otimes e_j) = \sum_j (Ae_j) \otimes e_j.$$

(iii) The space $\mathcal{L}(\mathcal{H}, \mathcal{K})$ is a Hilbert space with inner product $\langle A, B \rangle = \operatorname{tr} A^*B$. The set E_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n$, is an orthonormal basis for this space. Show that the map φ is a Hilbert space isomorphism; i.e., $\langle \varphi^{-1}(A), \varphi^{-1}(B) \rangle = \langle A, B \rangle$ for all A, B.

Corresponding facts about multilinear functionals and tensor products of several spaces are proved in the same way. We will use the notation $\otimes^k \mathcal{H}$ for the k-fold tensor product $\mathcal{H} \otimes \mathcal{H} \otimes \cdots \otimes \mathcal{H}$.

Tensor products of linear operators are defined as follows. We first define $A \otimes B$ on elementary tensors by putting $(A \otimes B)(x \otimes y) = Ax \otimes By$. We then extend this definition linearly to all linear combinations of elementary tensors, i.e., to all of $\mathcal{H} \otimes \mathcal{H}$. This extension involves no inconsistency.

It is obvious that $(A \otimes B)(C \otimes D) = AC \otimes BD$, that the identity on $\mathcal{H} \otimes \mathcal{H}$ is given by $I \otimes I$, and that if A and B are invertible, then so is $A \otimes B$ and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. A one-line verification shows that $(A \otimes B)^* = A^* \otimes B^*$. It follows that $A \otimes B$ is Hermitian if (but not only if) A and B are Hermitian; $A \otimes B$ is unitary if (but not only if) A and B are unitary; $A \otimes B$ is normal if (and only if) A and B are normal. (The trivial cases $A = 0$, or $B = 0$, must be excluded for the last assertion to be valid.)

Exercise I.4.5 *Suppose it is known that $\mathcal{M}$ is an invariant subspace for A. What invariant subspaces for $A \otimes A$ can be obtained from this information alone?*

For operators A, B on different spaces $\mathcal{H}$ and $\mathcal{K}$, the tensor product can be defined in the same way as above. This gives an operator $A \otimes B$ on $\mathcal{H} \otimes \mathcal{K}$. Many of the assertions made earlier for the case $\mathcal{H} = \mathcal{K}$ remain true in this situation.

Exercise I.4.6 *Let A and B be two matrices (not necessarily of the same size). Relative to the lexicographically ordered basis on the space of tensors, the matrix for $A \otimes B$ can be written in block form as follows: if $A = (a_{ij})$, then*

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \cdots & \cdots & \cdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}.$$

Especially important are the operators $A \otimes A \otimes \cdots \otimes A$, which are k-fold tensor products of an operator $A \in \mathcal{L}(\mathcal{H})$. Such a product will be written more briefly as $A^{\otimes k}$ or $\otimes^k A$. This is an operator on the n^k-dimensional space $\otimes^k \mathcal{H}$.

Some of the easily proved and frequently used properties of these products are summarised below:

1. $(\otimes^k A)(\otimes^k B) = \otimes^k (AB)$.

2. $\left(\otimes^k A\right)^{-1} = \otimes^k A^{-1}$ when either inverse exists.

3. $\left(\otimes^k A\right)^* = \otimes^k A^*$.

4. If A is Hermitian, unitary, normal or positive, then so is $\otimes^k A$.

5. If $\alpha_1, \ldots, \alpha_k$ (not necessarily distinct) are eigenvalues of A with eigenvectors $u_1, \ldots, u_k$, respectively, then $\alpha_1 \cdots \alpha_k$ is an eigenvalue of $\otimes^k A$ and $u_1 \otimes \cdots \otimes u_k$ is an eigenvector for it.

6. If $s_{i_1}, \ldots, s_{i_k}$ (not necessarily distinct) are singular values of A, then $s_{i_1} \cdots s_{i_k}$ is a singular value of $\otimes^k A$.

7. $\|\otimes^k A\| = \|A\|^k$.

The reader should formulate and prove analogous statements for tensor products $A_1 \otimes A_2 \otimes \cdots \otimes A_k$ of different operators.

I.5 Symmetry Classes

In the space $\otimes^k\mathcal{H}$ there are two especially important subspaces (for non-trivial cases, $k > 1$ and $n > 1$).

The **antisymmetric tensor product** of vectors $x_1, \ldots, x_k$ in $\mathcal{H}$ is defined as

$$x_1 \wedge \cdots \wedge x_k = (k!)^{-1/2} \sum_\sigma \varepsilon_\sigma x_{\sigma(1)} \otimes \cdots \otimes x_{\sigma(k)},$$

where σ runs over all permutations of the k indices and ε_σ is ± 1, depending on whether σ is an even or an odd permutation. (ε_σ is called the **signature** of σ.) The factor $(k!)^{-1/2}$ is chosen so that if x_j are orthonormal, then $x_1 \wedge \cdots \wedge x_k$ is a unit vector. The antisymmetry of this product means that

$$x_1 \wedge \cdots \wedge x_i \wedge \cdots \wedge x_j \wedge \cdots \wedge x_k = -x_1 \wedge \cdots \wedge x_j \wedge \cdots \wedge x_i \wedge \cdots \wedge x_k,$$

i.e., interchanging the position of any two of the factors in the product amounts to a change of sign. In particular, $x_1 \wedge \cdots \wedge x_k = 0$ if any two of the factors are equal.

The span of all antisymmetric tensors $x_1 \wedge \cdots \wedge x_k$ in $\otimes^k\mathcal{H}$ is denoted by $\wedge^k\mathcal{H}$. This is called the kth **antisymmetric tensor product** (or **tensor power**) of $\mathcal{H}$.

Given an orthonormal basis $(e_1, \ldots, e_n)$ in $\mathcal{H}$, there is a standard way of constructing an orthonormal basis in $\wedge^k\mathcal{H}$. Let $Q_{k,n}$ denote the set of all strictly increasing k-tuples chosen from $\{1, 2, \ldots, n\}$; i.e., $\mathcal{I} \in Q_{k,n}$ if and only if $\mathcal{I} = (i_1, i_2, \ldots, i_k)$, where $1 \le i_1 < i_2 < \cdots < i_k \le n$. For such an $\mathcal{I}$ let $e_\mathcal{I} = e_{i_1} \wedge \cdots \wedge e_{i_k}$. Then, $\{e_\mathcal{I} : \mathcal{I} \in Q_{k,n}\}$ gives an orthonormal basis of $\wedge^k\mathcal{H}$. Such $\mathcal{I}$ are sometimes called **multi-indices** . It is conventional to order them lexicographically. Note that the cardinality of $Q_{k,n}$, and hence the dimensionality of $\wedge^k\mathcal{H}$, is $\binom{n}{k}$.

If in particular $k = n$, the space $\wedge^k\mathcal{H}$ is 1-dimensional. This plays a special role later on. When $k > n$ the space $\wedge^k\mathcal{H}$ is $\{0\}$.

Exercise I.5.1 *Show that the inner product* $\langle x_1 \wedge \cdots \wedge x_k, y_1 \wedge \cdots \wedge y_k\rangle$ *is equal to the determinant of the* $k \times k$ *matrix* $(\langle x_i, y_j\rangle)$.

The **symmetric tensor product** of $x_1, \ldots, x_k$ is defined as

$$x_1 \vee \cdots \vee x_k = (k!)^{-1/2} \sum_\sigma x_{\sigma(1)} \otimes \cdots \otimes x_{\sigma(k)},$$

where σ, as before, runs over all permutations of the k indices. The linear span of all these vectors comprises the subspace $\vee^k\mathcal{H}$ of $\otimes^k\mathcal{H}$. This is called the kth **symmetric tensor power** of $\mathcal{H}$.

Let $G_{k,n}$ denote the set of all non-decreasing k-tuples chosen from $\{1, 2, \ldots, n\}$; i.e., $\mathcal{I} \in G_{k,n}$ if and only if $\mathcal{I} = (i_1, \ldots, i_k)$, where $1 \le i_1 \le i_2 \cdots \le i_k \le n$. If such an $\mathcal{I}$ consists of ℓ distinct indices $i_1, \ldots, i_\ell$ with multiplicities $m_1, \ldots, m_\ell$, respectively, put $m(\mathcal{I}) = m_1!m_2!\cdots m_\ell!$. Given

an orthonormal basis $(e_1, \ldots, e_n)$ of $\mathcal{H}$ define, for every $\mathcal{I} \in G_{k,n}$, $e_{\mathcal{I}} = e_{i_1} \vee e_{i_2} \vee \cdots \vee e_{i_k}$. Then, the set $\{m(\mathcal{I})^{-1/2} e_{\mathcal{I}} : \mathcal{I} \in G_{k,n}\}$ is an orthonormal basis in $\vee^k \mathcal{H}$. Again, it is conventional to order these multi-indices lexicographically. The cardinality of the set $G_{k,n}$, and hence the dimensionality of the space $\vee^k \mathcal{H}$, is $\binom{n+k-1}{k}$.

Notice that the expressions for the basis in $\wedge^k \mathcal{H}$ are simpler because $m(\mathcal{I}) = 1$ for $\mathcal{I} \in Q_{k,n}$.

Exercise I.5.2 *The elementary tensors $x \otimes \cdots \otimes x$, with all factors equal, are all in the subspace $\vee^k \mathcal{H}$. Do they span it?*

Exercise I.5.3 *Let $\mathcal{M}$ be a p-dimensional subspace of $\mathcal{H}$ and $\mathcal{N}$ its orthogonal complement. Choosing j vectors from $\mathcal{M}$ and $k - j$ vectors from $\mathcal{N}$ and forming the linear span of the antisymmetric tensor products of all such vectors, we get different subspaces of $\wedge^k \mathcal{H}$; for example, one of those is $\wedge^k \mathcal{M}$. Determine all the subspaces thus obtained and their dimensionalities. Do the same for $\vee^k \mathcal{H}$.*

Exercise I.5.4 *If* $\dim \mathcal{H} = 3$, *then* $\dim \otimes^3 \mathcal{H} = 27$, $\dim \wedge^3 \mathcal{H} = 1$ *and* $\dim \vee^3 \mathcal{H} = 10$. *In terms of an orthonormal basis of $\mathcal{H}$, write an element of* $(\wedge^3 \mathcal{H} \oplus \vee^3 \mathcal{H})^{\perp}$.

The **permanent** of a matrix $A = (a_{ij})$ is defined as

$$\text{per}\, A = \sum_{\sigma} a_{1\sigma(1)} \cdots a_{n\sigma(n)},$$

where σ varies over all permutations on n symbols. Note that, in contrast to the determinant, the permanent is not invariant under similarities. Thus, matrices of the same operator relative to different bases may have different permanents.

Exercise I.5.5 *Show that the inner product $\langle x_1 \vee \cdots \vee x_k, y_1 \vee \cdots \vee y_k \rangle$ is equal to the permanent of the $k \times k$ matrix $(\langle x_i, y_j \rangle)$.*

The spaces $\wedge^k \mathcal{H}$ and $\vee^k \mathcal{H}$ are also referred to as "symmetry classes" of tensors – there are other such classes in $\otimes^k \mathcal{H}$. Another way to look at them is as the ranges of the respective symmetry operators. Define $P_\wedge$ and $P_\vee$ as linear operators on $\otimes^k \mathcal{H}$ by first defining them on the elementary tensors as

$$P_\wedge(x_1 \otimes \cdots \otimes x_k) = (k!)^{-1/2} x_1 \wedge \cdots \wedge x_k$$

$$P_\vee(x_1 \otimes \cdots \otimes x_k) = (k!)^{-1/2} x_1 \vee \cdots \vee x_k$$

and extending them by linearity to the whole space. (Again it should be verified that this can be done consistently.) The constant factor in the above definitions has been chosen so that both these operators are idempotent. They are also Hermitian. The ranges of these orthoprojectors are $\wedge^k \mathcal{H}$ and $\vee^k \mathcal{H}$, respectively.

If $A \in \mathcal{L}(\mathcal{H})$, then $Ax_1 \wedge \cdots \wedge Ax_k$ lies in $\wedge^k\mathcal{H}$ for all $x_1, \ldots, x_k$ in $\mathcal{H}$. Using this, one sees that the space $\wedge^k\mathcal{H}$ is invariant under the operator $\otimes^k A$. The restriction of $\otimes^k A$ to this invariant subspace is denoted by $\wedge^k A$ or $A^{\wedge k}$. This is called the kth **antisymmetric tensor power** or the kth **Grassmann power** of A. We could have also defined it by first defining it on the elementary antisymmetric tensors $x_1 \wedge \cdots \wedge x_k$ as

$$\wedge^k A(x_1 \wedge \cdots \wedge x_k) = Ax_1 \wedge \cdots \wedge Ax_k$$

and then extending it linearly to the span $\wedge^k\mathcal{H}$ of these tensors.

Exercise I.5.6 *Let A be a nilpotent operator. Show how to obtain, from a Jordan basis for A, a Jordan basis for $\wedge^2 A$.*

The space $\vee^k\mathcal{H}$ is also invariant under the operator $\otimes^k A$. The restriction of $\otimes^k A$ to this invariant subspace is written as $\vee^k A$ or $A^{\vee k}$ and called the kth **symmetric tensor power** of A.

Some essential and simple properties of these operators are summarised below:

1. $(\wedge^k A)(\wedge^k B) = \wedge^k(AB), \quad (\vee^k A)(\vee^k B) = \vee^k(AB)$.

2. $(\wedge^k A)^* = \wedge^k A^*, \quad (\vee^k A)^* = \vee^k A^*$.

3. $(\wedge^k A)^{-1} = \wedge^k A^{-1}, \quad (\vee^k A)^{-1} = \vee^k A^{-1}$.

4. If A is Hermitian, unitary, normal or positive, then so are $\wedge^k A$ and $\vee^k A$.

5. If $\alpha_1, \ldots, \alpha_k$ are eigenvalues of A (not necessarily distinct) belonging to eigenvectors $u_1, \ldots, u_k$, respectively, then $\alpha_1 \cdots \alpha_k$ is an eigenvalue of $\vee^k A$ belonging to eigenvector $u_1 \vee \cdots \vee u_k$; if in addition the vectors u_j are linearly independent, then $\alpha_1 \cdots \alpha_k$ is an eigenvalue of $\wedge^k A$ belonging to eigenvector $u_1 \wedge \cdots \wedge u_k$.

6. If $s_1, \ldots, s_n$ are the singular values of A, then the singular values of $\wedge^k A$ are $s_{i_1} \cdots s_{i_k}$, where $(i_1, \ldots, i_k)$ vary over $Q_{k,n}$; the singular values of $\vee^k A$ are $s_{i_1} \cdots s_{i_k}$, where $(i_i, \ldots, i_k)$, vary over $G_{k,n}$.

7. $\text{tr}\wedge^k A$ is the kth elementary symmetric polynomial in the eigenvalues of A; $\text{tr}\vee^k A$ is the kth complete symmetric polynomial in the eigenvalues of A.

(These polynomials are defined as follows. Given any n-tuple $(\alpha_1, \ldots, \alpha_n)$ of numbers or other commuting objects, the kth **elementary symmetric polynomial** in them is the sum of all terms $\alpha_{i_1}\alpha_{i_2} \cdots \alpha_{i_k}$ for $(i_1, i_2, \ldots, i_k)$ in $Q_{k,n}$; the kth **complete symmetric polynomial** is the sum of all terms $\alpha_{i_1}\alpha_{i_2} \cdots \alpha_{i_k}$ for $(i_1, i_2, \ldots, i_k)$ in $G_{k,n}$.)

For $A \in \mathcal{L}(\mathcal{H})$, consider the operator $A \otimes I \otimes \cdots \otimes I + I \otimes A \otimes I \cdots \otimes I + \cdots + I \otimes I \otimes \cdots \otimes A$. (There are k summands, each of which is a product

of k factors.) The eigenvalues of this operator on $\otimes^k \mathcal{H}$ are sums of eigenvalues of A. Both the spaces $\wedge^k \mathcal{H}$ and $\vee^k \mathcal{H}$ are invariant under this operator. One pleasant way to see this is to regard this operator as the t-derivative at $t = 0$ of $\otimes^k (I + tA)$. The restriction of this operator to the space $\wedge^k \mathcal{H}$ will be of particular interest to us; we will write this restriction as $A^{[k]}$. If $u_1, \ldots, u_k$ are linearly independent eigenvectors of A belonging to eigenvalues $\alpha_1, \ldots, \alpha_k$, then $u_1 \wedge \cdots \wedge u_k$ is an eigenvector of $A^{[k]}$ belonging to eigenvalue $\alpha_1 + \cdots + \alpha_k$.

Now, fixing an orthonormal basis $(e_1, \ldots, e_n)$ of $\mathcal{H}$, identify A with its matrix (a_{ij}). We want to find the matrix representations of $\wedge^k A$ and $\vee^k A$ relative to the standard bases constructed earlier.

The basis of $\wedge^k \mathcal{H}$ we are using is $e_{\mathcal{I}}, \mathcal{I} \in Q_{k,n}$. The $(\mathcal{I}, \mathcal{J})$-entry of $\wedge^k A$ is $\langle e_{\mathcal{I}}, (\wedge^k A) e_{\mathcal{J}} \rangle$. One may verify that this is equal to a subdeterminant of A. Namely, let $A[\mathcal{I}|\mathcal{J}]$ denote the $k \times k$ matrix obtained from A by expunging all its entries a_{ij} except those for which $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Then, the $(\mathcal{I}, \mathcal{J})$-entry of $\wedge^k A$ is equal to det $A[\mathcal{I}|\mathcal{J}]$.

The special case $k = n$ leads to the 1-dimensional space $\wedge^n \mathcal{H}$. The operator $\wedge^n A$ on this space is just the operator of multiplication by the number det A. We can thus think of det A as being equal to $\wedge^n A$.

The basis of $\vee^k \mathcal{H}$ we are using is $m(\mathcal{I})^{-1/2} e_{\mathcal{I}}$, $\mathcal{I} \in G_{k,n}$. The $(\mathcal{I}, \mathcal{J})$-entry of the matrix $\vee^k A$ can be computed as before, and the result is somewhat similar to that for $\wedge^k A$. For $\mathcal{I} = (i_1, \ldots, i_k)$ and $\mathcal{J} = (j_1, \ldots, j_k)$ in $G_{k,n}$, let $A[\mathcal{I}|\mathcal{J}]$ now denote the $k \times k$ matrix whose (r, s)-entry is the (i_r, j_s) - entry of A. Since repetitions of indices are allowed in $\mathcal{I}$ and $\mathcal{J}$, this is not a submatrix of A this time. One verifies that the $(\mathcal{I}, \mathcal{J})$-entry of $\vee^k A$ is $(m(\mathcal{I}) m(\mathcal{J}))^{-1/2}$ per $A[\mathcal{I}|\mathcal{J}]$.

In particular, per A is one of the diagonal entries of $\vee^n A$: the $(\mathcal{I}, \mathcal{I})$-entry for $\mathcal{I} = (1, 2, \ldots, n)$.

Exercise I.5.7 *Prove that for any vectors* $u_1, \ldots, u_k, v_1, \ldots, v_k$ *we have*

$$\begin{aligned} |\det(\langle u_i, v_j \rangle)|^2 &\leq \det(\langle u_i, u_j \rangle) \det(\langle v_i, v_j \rangle), \\ |\mathrm{per}(\langle u_i, v_j \rangle)|^2 &\leq \mathrm{per}(\langle u_i, u_j \rangle) \mathrm{per}(\langle v_i, v_j \rangle). \end{aligned}$$

Exercise I.5.8 *Prove that for any two matrices* A, B *we have*

$$|\mathrm{per}(AB)|^2 \leq \mathrm{per}(AA^*) \mathrm{per}(B^* B).$$

(The corresponding relation for determinants is an easy equality.)

Exercise I.5.9 *(***Schur's Theorem***) If* A *is positive, then*

$$\mathrm{per}\, A \geq \det A.$$

[Hint: Using Exercise I.2.2 write $A = T^* T$ *for an upper triangular* T*. Then use the preceding exercise cleverly.]*

We have observed earlier that for any vectors $x_1, \ldots, x_k$ in $\mathcal{H}$ we have

$$\det(\langle x_i, x_j \rangle) = \|x_1 \wedge \cdots \wedge x_k\|^2.$$

When $\mathcal{H} = \mathbb{R}^n$, this determinant is also the square of the k-dimensional volume of the parallelepiped having $x_1, \ldots, x_k$ as its sides. To see this, note that neither the determinant nor the volume in question is altered if we add to any of these vectors a linear combination of the others. Performing such operations successively, we can reach an orthogonal set of vectors, some of which might be zero. In this case it is obvious that the determinant is equal to the square of the volume; hence that was true initially too.

Given any k-tuple $X = (x_1, \ldots, x_k)$, the matrix $(\langle x_i, x_j \rangle) = X^*X$ is called the **Gram matrix** of the vectors x_j; its determinant is called their **Gram determinant**.

Exercise I.5.10 *Every $k \times k$ positive matrix $A = (a_{ij})$ can be realised as a Gram matrix, i.e., vectors $x_j, 1 \leq j \leq k$, can be found so that $a_{ij} = \langle x_i, x_j \rangle$ for all i, j.*

I.6 Problems

Problem I.6.1. Given a basis $U = (u_1, \ldots, u_n)$, not necessarily orthonormal, in $\mathcal{H}$, how would you compute the biorthogonal basis $(v_1, \ldots, v_n)$? Find a formula that expresses $\langle v_j, x \rangle$ for each $x \in \mathcal{H}$ and $j = 1, 2, \ldots, k$ in terms of Gram matrices.

Problem I.6.2. A proof of the Toeplitz-Hausdorff Theorem is outlined below. Fill in the details.

Note that $W(A) = \{\langle x, Ax \rangle : \|x\| = 1\} = \{\operatorname{tr} Axx^* : x^*x = 1\}$. It is enough to consider the special case $\dim \mathcal{H} = 2$. In higher dimensions, this special case can be used to show that if x, y are any two vectors, then any point on the line segment joining $\langle x, Ax \rangle$ and $\langle y, Ay \rangle$ can be represented as $\langle z, Az \rangle$, where z is a vector in the linear span of x and y. Now, on the space of 2×2 Hermitian matrices consider the linear map $\Phi(T) = \operatorname{tr} AT$. This is a real linear map from a space of 4 real dimensions (the 2×2 Hermitian matrices) to a space of 2 real dimensions (the complex plane). We want to prove that Φ maps the set of 1-dimensional orthoprojectors xx^* onto a convex set. The set of these projectors in matrix form is

$$\begin{pmatrix} \cos t \\ e^{-iw} \sin t \end{pmatrix} (\cos t \;\; e^{iw} \sin t) = \frac{1}{2} + \frac{1}{2} \begin{pmatrix} \cos 2t & e^{iw} \sin 2t \\ e^{-iw} \sin 2t & -\cos 2t \end{pmatrix}.$$

This is a 2-sphere centred at $\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$ and having radius $1/\sqrt{2}$ in the Frobenius norm. The image of a 2-sphere under a linear map with range in $\mathbb{R}^2$

must be either an ellipse with interior, or a line segment, or a point; in any case, a convex set.

Problem I.6.3. By the remarks in Section 5, vectors $x_1, \ldots, x_k$ are linearly dependent if and only if $x_1 \wedge \cdots \wedge x_k = 0$. This relationship between linear dependence and the antisymmetric tensor product goes further. Two sets $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_k\}$ of linearly independent vectors have the same linear span if and only if $x_1 \wedge \cdots \wedge x_k = c y_1 \wedge \cdots \wedge y_k$ for some constant c. Thus, there is a one-to-one correspondence between k-dimensional subspaces of a vector space W and 1-dimensional subspaces of $\wedge^k W$ generated by elementary tensors $x_1 \wedge \cdots \wedge x_k$.

Problem I.6.4. How large must $\dim W$ be in order that there exist some element of $\wedge^2 W$ which is not elementary?

Problem I.6.5. Every vector w of W induces a linear operator T_w from $\wedge^k W$ to $\wedge^{k+1} W$ as follows. T_w is defined on elementary tensors as $T_w(v_1 \wedge \cdots \wedge v_k) = v_1 \wedge \cdots \wedge v_k \wedge w$, and then extended linearly to all of $\wedge^k W$. It is, then, natural to write $T_w(x) = x \wedge w$ for any $x \in \wedge^k W$. Show that a nonzero vector x in $\wedge^k W$ is elementary if and only if the space $\{w \in W : x \wedge w = 0\}$ is k-dimensional.

(When W is a Hilbert space, the operators T_w are called **creation operators** and their adjoints are called **annihilation operators** in the physics literature.)

Problem I.6.6. (The n-dimensional Pythagorean Theorem) Let $x_1, \ldots, x_n$ be orthogonal vectors in $\mathbb{R}^n$. Consider the n-dimensional simplex S with vertices $0, x_1, \ldots, x_n$. Think of the $(n-1)$-dimensional simplex with vertices $x_1, \ldots, x_n$ as the "hypotenuse" of S and the remaining $(n-1)$-dimensional faces of S as its "legs". By the remarks in Section 5, the k-dimensional volume of the simplex formed by any k points $y_1, \ldots, y_k$ together with the origin is $(k!)^{-1} \|y_1 \wedge \cdots \wedge y_k\|$. The volume of a simplex not having 0 as a vertex can be found by translating it. Use this to prove that the square of the volume of the hypotenuse of S is the sum of the squares of the volumes of the n legs.

Problem I.6.7. (i) Let $Q_\wedge$ be the inclusion map from $\wedge^k \mathcal{H}$ into $\otimes^k \mathcal{H}$ (so that $Q_\wedge^*$ equals the projection $P_\wedge$ defined earlier) and let $Q_\vee$ be the inclusion map from $\vee^k \mathcal{H}$ into $\otimes^k \mathcal{H}$. Then, for any $A \in \mathcal{L}(\mathcal{H})$

$$\wedge^k A = P_\wedge (\otimes^k A) Q_\wedge,$$

$$\vee^k A = P_\vee (\otimes^k A) Q_\vee.$$

(ii) $\| \wedge^k A\| \leq \|A\|^k, \quad \| \vee^k A\| \leq \|A\|^k.$

(iii) $|\det A| \leq \|A\|^n, \quad |\text{per} A| \leq \|A\|^n.$

Problem I.6.8. For an invertible operator A obtain a relationship between $A^{-1}, \wedge^n A$, and $\wedge^{n-1} A$.

Problem I.6.9. (i) Let $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ be two orthonormal bases in $\mathcal{H}$. Show that

$$|\langle e_2 \wedge \cdots \wedge e_n, f_2 \wedge \cdots \wedge f_n \rangle|^2 = |\langle e_1, f_1 \rangle|^2.$$

(ii) Let P and Q be orthogonal projections in $\mathcal{H}$, each of rank $n-1$. Let x, y be unit vectors such that $Px = Qy = 0$. Show that

$$\wedge^{n-1}(PQP) = |\langle x, y \rangle|^2 \ \wedge^{n-1} P.$$

Problem I.6.10. If the characteristic polynomial of A is written as

$$t^n + a_1 t^{n-1} + \cdots + a_n,$$

then the coefficient a_k is the sum of all $k \times k$ principal minors of A. This is equal to $\operatorname{tr} \wedge^k A$.

Problem I.6.11. (i) For any $A, B \in \mathcal{L}(\mathcal{H})$ we have

$$\otimes^k A - \otimes^k B = \sum_{j=1}^{k} C_j \ ,$$

where

$$C_j = (\otimes^{k-j} A) \otimes (A - B) \otimes (\otimes^{j-1} B).$$

Hence,

$$\| \otimes^k A - \otimes^k B \| \leq k M^{k-1} \|A - B\|,$$

where $M = \max(\|A\|, \|B\|)$.

(ii) The norms of $\wedge^k A - \wedge^k B$ and $\vee^k A - \vee^k B$ are therefore also bounded by $kM^{k-1}\|A - B\|$.

(iii) For $n \times n$ matrices A, B,

$$|\det A - \det B| \leq n M^{n-1} \|A - B\|,$$

$$|\operatorname{per} A - \operatorname{per} B| \leq n M^{n-1} \|A - B\|.$$

(iv) The example $A = \alpha I, B = (\alpha + \varepsilon) I$ for small ε shows that these inequalities are sometimes sharp. When $\|A\|$ and $\|B\|$ are far apart, find a simple improvement on them.

(v) If A, B are $n \times n$ matrices with characteristic polynomials

$$t^n + a_1 t^{n-1} + \cdots + a_n,$$

$$t^n + b_1 t^{n-1} + \cdots + b_n,$$

respectively, then

$$|a_k - b_k| \leq k \binom{n}{k} M^{k-1} \|A - B\|,$$

where $M = \max(\|A\|, \|B\|)$.

Problem I.6.12. Let A, B be positive operators with $A \geq B$ (i.e., $A - B$ is positive). Show that

$$\begin{aligned} \otimes^k A &\geq \otimes^k B, \\ \wedge^k A &\geq \wedge^k B, \\ \vee^k A &\geq \vee^k B, \\ \det A &\geq \det B, \\ \operatorname{per} A &\geq \operatorname{per} B. \end{aligned}$$

Problem I.6.13. The **Schur product** or the **Hadamard product** of two matrices A and B is defined to be the matrix $A \circ B$ whose (i,j)-entry is $a_{ij}b_{ij}$. Show that this is a principal submatrix of $A \otimes B$, and derive from this fact two significant properties:

(i) $\|A \circ B\| \leq \|A\| \, \|B\|$ for all A, B.

(ii) If A, B are positive, then so is $A \circ B$. (This is called **Schur's Theorem.**)

Problem I.6.14. (i) Let $A = (a_{ij})$ be an $n \times n$ positive matrix. Let

$$\begin{aligned} r_i &= \sum_{j=1}^{n} a_{ij}, \quad 1 \leq i \leq n, \\ s &= \sum_{i,j} a_{ij}. \end{aligned}$$

Show that

$$s^n \operatorname{per} A \geq n! \prod_{i=1}^{n} |r_i|^2.$$

[Hint: Represent A as the Gram matrix of some vectors $x_1, \ldots, x_n$ as in Exercise I.5.10. Let $u = s^{-1/2}(x_1 + \cdots + x_n)$. Consider the vectors $u \vee u \vee \cdots \vee u$ and $x_1 \vee \cdots \vee x_n$, and use the Cauchy-Schwarz inequality.]

(ii) Show that equality holds in the above inequality if and only if either A has rank 1 or A has a row of zeroes.

(iii) If in addition all a_{ij} are nonnegative and all $r_i = 1$ (so that the matrix A is doubly stochastic as well as positive semidefinite), then

$$\operatorname{per} A \geq \frac{n!}{n^n}.$$

Here equality holds if and only if $a_{ij} = \frac{1}{n}$ for all i, j.

Problem I.6.15. Let A be Hermitian with eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$. In Exercise I.2.7 we noted that

$$\alpha_1 = \max\{\langle x, Ax\rangle : \|x\| = 1\},$$

$$\alpha_n = \min\{\langle x, Ax\rangle : \|x\| = 1\}.$$

Using these relations and tensor products, we can deduce some other extremal representations:

(i) For every $k = 1, 2, \ldots, n$,

$$\sum_{j=1}^{k} \alpha_j = \max \sum_{j=1}^{k} \langle x_j, Ax_j\rangle,$$

$$\sum_{j=n-k+1}^{n} \alpha_j = \min \sum_{j=1}^{k} \langle x_j, Ax_j\rangle,$$

where the maximum and the minimum are taken over all choices of orthonormal k-tuples $(x_1, \ldots, x_k)$ in $\mathcal{H}$. The first statement is referred to as **Ky Fan's Maximum Principle**. It will reappear in Chapter II (with a different proof) and subsequently.

(ii) If A is positive, then for every $k = 1, 2, \ldots, n$,

$$\prod_{j=n-k+1}^{n} \alpha_j = \min \prod_{j=1}^{k} \langle x_j, Ax_j\rangle,$$

where the minimum is taken over all choices of orthonormal k-tuples $(x_1, \ldots, x_k)$ in $\mathcal{H}$.

[Hint: You may need to use the *Hadamard Determinant Theorem*, which says that the determinant of a positive matrix is bounded above by the product of its diagonal entries. This is also proved in Chapter II.]

(ii) If A is positive, then for every $\mathcal{I} \in Q_{k,n}$

$$\prod_{j=n-k+1}^{n} \alpha_j \leq \det A[\mathcal{I}|\mathcal{I}] \leq \prod_{j=1}^{k} \alpha_j.$$

Problem I.6.16. Let A be any $n \times n$ matrix with eigenvalues $\alpha_1, \ldots, \alpha_n$. Show that

$$\left|\alpha_j - \frac{\operatorname{tr} A}{n}\right| \leq \left[\frac{n-1}{n}\left(\|A\|_2^2 - \frac{|\operatorname{tr} A|^2}{n}\right)\right]^{1/2}$$

for all $j = 1, 2, \ldots, n$. (Results such as this are interesting because they give some information about the location of the eigenvalues of a matrix in

terms of more easily computable functions like the Frobenius norm $\|A\|_2$ and the trace. We will see several such statements later.)

[Hint: First prove that if $x = (x_1, \dots, x_n)$ is a vector with $x_1 + \cdots + x_n = 0$, then

$$\max |x_j| \leq \left(\frac{n-1}{n}\right)^{1/2} \|x\| \ .]$$

Problem I.6.17. (i) Let z_1, z_2, z_3 be three points on the unit circle. Then, the numerical range of an operator A is contained in the triangle with vertices z_1, z_2, z_3 if and only if A can be expressed as $A = z_1A_1 + z_2A_2 + z_3A_3$, where A_1, A_2, A_3 are positive operators with $A_1 + A_2 + A_3 = I$.

[Hint: It is easy to see that if A is a sum of this form, then $W(A)$ is contained in the given triangle. The converse needs some work to prove. Let z be any point in the given triangle. Then, one can find $\alpha_1, \alpha_2, \alpha_3$ such that $\alpha_j \geq 0, \alpha_1 + \alpha_2 + \alpha_3 = 1$ and $z = \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3$. These are the "barycentric coordinates" of z and can be obtained as follows. Let $\gamma = \operatorname{Im}(\bar{z}_1 z_2 + \bar{z}_2 z_3 + \bar{z}_3 z_1)$. Then, for $j = 1, 2, 3$,

$$\alpha_j = \operatorname{Im} \frac{(z - z_{j+1})(\bar{z}_{j+2} - \bar{z}_{j+1})}{\gamma},$$

where the subscript indices are counted modulo 3. Put

$$A_j = \operatorname{Im} \frac{(A - z_{j+1}I)(\bar{z}_{j+2} - \bar{z}_{j+1})}{\gamma}.$$

Then, A_j have the required properties.]

(ii) Let $W(A)$ be contained in a triangle with vertices z_1, z_2, z_3 lying on the unit circle. Then, choosing A_1, A_2, A_3 as above, write

$$\begin{pmatrix} I & A^* \\ A & I \end{pmatrix} = \sum_{j=1}^{3} \begin{pmatrix} A_j & \bar{z}_j A_j \\ z_j A_j & A_j \end{pmatrix} = \sum_{j=1}^{3} A_j \otimes \begin{pmatrix} 1 & \bar{z}_j \\ z_j & 1 \end{pmatrix}.$$

This, being a sum of three positive matrices, is positive. Hence, by Proposition I.3.5 A is a contraction.

(iii) If $W(A)$ is contained in a triangle with vertices z_1, z_2, z_3, then $\|A\| \leq \max |z_j|$. This is **Mirman's Theorem.**

Problem I.6.18. If an operator T has the Cartesian decomposition $T = A + iB$ with A and B positive, then

$$\|T\|^2 \leq \|A\|^2 + \|B\|^2.$$

Show that, if A or B is not positive then this need not be true.

[Hint: To prove the above inequality note that $W(T)$ is contained in a rectangle in the first quadrant. Find a suitable triangle that contains it and use Mirman's Theorem.]

I.7 Notes and References

Standard references on linear algebra and matrix theory include P.R. Halmos, *Finite-Dimensional Vector Spaces*, Van Nostrand, 1958; F.R. Gantmacher, *Matrix Theory*, 2 volumes, Chelsea, 1959 and K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed., Prentice Hall, 1971. A recent work is R.A. Horn and C.R. Johnson, in two volumes, *Matrix Analysis* and *Topics in Matrix Analysis*, Cambridge University Press, 1985 and 1990. For more of multilinear algebra, see W. Greub, *Multilinear Algebra*, 2nd ed., Springer-Verlag, 1978, and M. Marcus, *Finite-Dimensional Multilinear Algebra*, 2 volumes, Marcel Dekker, 1973 and 1975. A brief treatment that covers all the basic results may be found in M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Prindle, Weber and Schmidt, 1964, reprinted by Dover in 1992.

Though not as important as the determinant, the permanent of a matrix is an interesting object with many uses in combinatorics, geometry, and physics. A book devoted entirely to it is H. Minc, *Permanents*, Addison-Wesley, 1978.

Apart from the symmetric and the antisymmetric tensors, there are other symmetry classes of tensors. Their study is related to the glorious subject of representations of finite groups. See J.P. Serre, *Linear Representations of Finite Groups*, Springer-Verlag, 1977.

The result in Exercise I.3.6 is due to P.R. Halmos, and is the beginning of a subject called Dilation Theory. See Chapter 23 of P.R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, 1982.

The proof of the Toeplitz-Hausdorff Theorem in Problem I.6.2 is taken from C. Davis, *The Toeplitz-Hausdorff theorem explained*, Canad. Math. Bull., 14(1971) 245-246. For a different proof, see P.R. Halmos, *A Hilbert Space Problem Book*.

For relations between Grassmann spaces and geometry, as indicated in Problem I.6.3, see, for example, I.R. Porteous, *Topological Geometry*, Cambridge University Press, 1981. The simple proof of the Pythagorean Theorem in Problem I.6.6 is due to S. Ramanan.

Among the several papers in quantum physics, where ideas very close to those in Problems I.6.3 and I.6.5 are used effectively, is one by N.M. Hugenholtz and R.V. Kadison, *Automorphisms and quasi-free states of the CAR algebra*, Commun. Math. Phys., 43 (1975) 181-197.

Inequalities like the ones in Problem I.6.11 were first discovered in connection with perturbation theory of eigenvalues. This is summarised in R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman, 1987. The simple identity at the beginning of Problem I.6.11 was first used in this context in R. Bhatia and L. Elsner, *On the variation of permanents*, Linear and Multilinear Algebra, 27(1990) 105-110.

The results and the ideas of Problem I.6.14 are from M. Marcus and M. Newman, *Inequalities for the permanent function*, Ann. of Math., 75(1962)

47-62. In 1926, B.L. van der Waerden had conjectured that the inequality in part (iii) of Problem I.6.14 will hold for all doubly stochastic matrices. This conjecture was proved, in two separate papers in 1981, by G.P. Egorychev and D. Falikman. An expository account is given in J.H. van Lint, *The van der Waerden conjecture: two proofs in one year*, Math. Intelligencer, 4(1982)72-77.

The results of Problem I.6.15 are all due to Ky Fan, *On a theorem of Weyl concerning eigenvalues of linear transformations I,II*, Proc. Nat. Acad. Sci., U.S.A., 35(1949) 652-655, 36(1950)31-35, and *A minimum property of the eigenvalues of a Hermitian transformation*, Amer. Math. Monthly, 60(1953)48-50.

A special case of the inequality of Problem I.6.16 occurs in P. Tarazaga, *Eigenvalue estimates for symmetric matrices*, Linear Algebra and Appl., 135(1990) 171-179.

Mirman's Theorem is proved in B.A. Mirman, *Numerical range and norm of a linear operator*, Trudy Seminara po Funkcional' nomu Analizu, No. 10 (1968), pp. 51-55. The inequality of Problem I.6.18 is also noted there as a corollary. Our proof of Mirman's Theorem is taken from Y. Nakamura, *Numerical range and norm*, Math. Japonica, 27 (1982) 149-150.

II

Majorisation and Doubly Stochastic Matrices

Comparison of two vector quantities often leads to interesting inequalities that can be expressed succinctly as "majorisation" relations. There is an intimate relation between majorisation and doubly stochastic matrices. These topics are studied in detail here. We place special emphasis on majorisation relations between the eigenvalue n-tuples of two matrices. This will be a recurrent theme in the book.

II.1 Basic Notions

Let $x = (x_1, \ldots, x_n)$ be an element of $\mathbb{R}^n$. Let $x^\downarrow$ and $x^\uparrow$ be the vectors obtained by rearranging the coordinates of x in the decreasing and the increasing orders, respectively. Thus, if $x^\downarrow = (x_1^\downarrow, \ldots, x_n^\downarrow)$,, then $x_1^\downarrow \geq \cdots \geq x_n^\downarrow$. Similarly, if $x^\uparrow = (x_1^\uparrow, \ldots, x_n^\uparrow)$,, then $x_1^\uparrow \leq \cdots \leq x_n^\uparrow$. Note that

$$x_j^\uparrow = x_{n-j+1}^\downarrow, \qquad 1 \leq j \leq n. \tag{II.1}$$

Let $x, y \in \mathbb{R}^n$. We say that x is **majorised** by y, in symbols $x \prec y$, if

$$\sum_{j=1}^{k} x_j^\downarrow \leq \sum_{j=1}^{k} y_j^\downarrow, \qquad 1 \leq k \leq n, \tag{II.2}$$

and

$$\sum_{j=1}^{n} x_j^\downarrow = \sum_{j=1}^{n} y_j^\downarrow. \tag{II.3}$$

Example: If $x_i \geq 0$ and $\Sigma x_i = 1$, then

$$(\frac{1}{n}, \ldots, \frac{1}{n}) \prec (x_1, \ldots, x_n) \prec (1, 0, \ldots, 0).$$

The notion of majorisation occurs naturally in various contexts. For example, in physics, the relation $x \prec y$ is interpreted to mean that the vector x describes a "more chaotic" state than y. (Think of x_i as the probability of a system being found in state i.) Another example occurs in economics. If $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ denote incomes of individuals $1, 2, \ldots, n$, then $x \prec y$ would mean that there is a more equal distribution of incomes in the state x than in y. The above example illustrates this.

From (II.1) we have

$$\sum_{j=1}^{k} x_j^{\uparrow} = \sum_{j=1}^{n} x_j - \sum_{j=1}^{n-k} x_j^{\downarrow}.$$

Hence $x \prec y$ if and only if

$$\sum_{j=1}^{k} x_j^{\uparrow} \geq \sum_{j=1}^{k} y_j^{\uparrow}, \quad 1 \leq k \leq n \tag{II.4}$$

and

$$\sum_{j=1}^{n} x_j^{\uparrow} = \sum_{j=1}^{n} y_j^{\uparrow}. \tag{II.5}$$

Let e denote the vector $(1, 1, \ldots, 1)$, and for any subset I of $\{1, 2, \ldots, n\}$ let e_I denote the vector whose jth component is 1 if $j \in I$ and 0 if $j \notin I$. Given a vector $x \in \mathbb{R}^n$, let

$$\text{tr } x = \sum_{j=1}^{n} x_j = \langle x, e \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^n$. Note that

$$\sum_{j=1}^{k} x_j^{\downarrow} = \max_{|I|=k} \langle x, e_I \rangle,$$

where $|I|$ stands for the number of elements in the set I.

So, $x \prec y$ if and only if for each subset I of $\{1, 2, \ldots, n\}$ there exists a subset J with $|I| = |J|$ such that

$$\langle x, e_I \rangle \leq \langle y, e_J \rangle \tag{II.6}$$

and

$$\text{tr } x = \text{tr } y. \tag{II.7}$$

We say that x is **(weakly) submajorised** by y, in symbols $x \prec_w y$, if condition (II.2) is fulfilled.

Note that in the absence of (II.3), the conditions (II.2) and (II.4) are not equivalent. We say that x is **(weakly) supermajorised** by y, in symbols $x \prec^w y$, if condition (II.4) is fulfilled.

Exercise II.1.1 *(i) $x \prec y \Leftrightarrow x \prec_w y$ and $x \prec^w y$.*

(ii) If α is a positive real number, then

$$x \prec_w y \Rightarrow \alpha x \prec_w \alpha y,$$

$$x \prec^w y \Rightarrow \alpha x \prec^w \alpha y.$$

(iii) $x \prec_w y \Leftrightarrow -x \prec^w -y$.

(iv) For any real number α,

$$x \prec y \Rightarrow \alpha x \prec \alpha y.$$

Remark II.1.2 *The relations $\prec, \prec_w$, and $\prec^w$ are all reflexive and transitive. None of them, however, is a partial order. For example, if $x \prec y$ and $y \prec x$, we can only conclude that $x = Py$, where P is a permutation matrix. If we say that $x \sim y$ whenever $x = Py$ for some permutation matrix P, then $\sim$ defines an equivalence relation on $\mathbb{R}^n$. If we denote by $\mathbb{R}^n_{sym}$ the resulting quotient space, then $\prec$ defines a partial order on this space. This relation is also a partial order on the set $\{x \in \mathbb{R}^n : x_1 \geq \cdots \geq x_n\}$. These statements are true for the relations $\prec_w$ and $\prec^w$ as well.*

For $a, b \in \mathbb{R}$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For $x, y \in \mathbb{R}^n$, define

$$x \vee y = (x_1 \vee y_1, \ldots, x_n \vee y_n)$$

$$x \wedge y = (x_1 \wedge y_1, \ldots, x_n \wedge y_n).$$

Let

$$\begin{aligned} x^+ &= x \vee 0, \\ |x| &= x \vee (-x). \end{aligned}$$

In other words, x^+ is the vector obtained from x by replacing the negative coordinates by zeroes, and $|x|$ is the vector obtained by taking the absolute values of all coordinates.

With these notations we can prove the following characterisation of majorisation that does not involve rearrangements:

Theorem II.1.3 *Let $x, y \in \mathbb{R}^n$. Then,*

(i) $x \prec_w y$ if and only if for all $t \in \mathbb{R}$

$$\sum_{j=1}^{n} (x_j - t)^+ \leq \sum_{j=1}^{n} (y_j - t)^+. \tag{II.8}$$

(ii) $x \prec^w y$ *if and only if for all* $t \in \mathbb{R}$

$$\sum_{j=1}^{n}(t - x_j)^+ \leq \sum_{j=1}^{n}(t - y_j)^+. \tag{II.9}$$

(iii) $x \prec y$ *if and only if for all* $t \in \mathbb{R}$

$$\sum_{j=1}^{n}|x_j - t| \leq \sum_{j=1}^{n}|y_j - t|. \tag{II.10}$$

Proof. Let $x \prec_w y$. If $t > x_1^{\downarrow}$, then $(x_j - t)^+ = 0$ for all j, and hence (II.8) holds. Let $x_{k+1}^{\downarrow} \leq t \leq x_k^{\downarrow}$ for some $1 \leq k \leq n$, where, for convenience, $x_{n+1}^{\downarrow} = -\infty$. Then,

$$\begin{aligned}\sum_{j=1}^{n}(x_j - t)^+ &= \sum_{j=1}^{k}(x_j^{\downarrow} - t) = \sum_{j=1}^{k}x_j^{\downarrow} - kt \\ &\leq \sum_{j=1}^{k}y_j^{\downarrow} - kt \leq \sum_{j=1}^{k}(y_j^{\downarrow} - t)^+,\end{aligned}$$

and, hence, (II.8) holds.

To prove the converse, note that if $t = y_k^{\downarrow}$, then

$$\sum_{j=1}^{n}(y_j - t)^+ = \sum_{j=1}^{k}(y_j^{\downarrow} - t) = \sum_{j=1}^{k}y_j^{\downarrow} - kt.$$

But

$$\begin{aligned}\sum_{j=1}^{k}x_j^{\downarrow} - kt &= \sum_{j=1}^{k}(x_j^{\downarrow} - t) \leq \sum_{j=1}^{k}(x_j^{\downarrow} - t)^+ \\ &\leq \sum_{j=1}^{n}(x_j^{\downarrow} - t)^+ = \sum_{j=1}^{n}(x_j - t)^+.\end{aligned}$$

So, if (II.8) holds, then we must have

$$\sum_{j=1}^{k}x_j^{\downarrow} \leq \sum_{j=1}^{k}y_j^{\downarrow},$$

i.e., $x \prec_w y$.

This proves (i). The statements (ii) and (iii) have similar proofs. ■

Corollary II.1.4 *If* $x \prec y$ *in* $\mathbb{R}^n$ *and* $u \prec w$ *in* $\mathbb{R}^m$*, then* $(x, u) \prec (y, w)$ *in* $\mathbb{R}^{n+m}$*. In particular,* $x \prec y$ *if and only if* $(x, u) \prec (y, u)$ *for all* u*.*

An $n \times n$ matrix $A = (a_{ij})$ is called **doubly stochastic** if

$$a_{ij} \geq 0 \qquad \text{for all} \quad i, j, \tag{II.11}$$

$$\sum_{i=1}^{n} a_{ij} = 1 \qquad \text{for all} \quad j, \tag{II.12}$$

$$\sum_{j=1}^{n} a_{ij} = 1 \qquad \text{for all} \quad i. \tag{II.13}$$

Exercise II.1.5 *A linear map A on $\mathbb{C}^n$ is called* **positivity-preserving** *if it carries vectors with nonnegative coordinates to vectors with nonnegative coordinates. It is called* **trace-preserving** *if* tr $Ax=$ tr x *for all x. It is called* **unital** *if $Ae = e$. Show that a matrix A is doubly stochastic if and only if the linear operator A is positivity-preserving, trace-preserving and unital. Show that A is trace-preserving if and only if its adjoint A^* is unital.*

Exercise II.1.6 *(i) The class of $n \times n$ doubly stochastic matrices is a convex set and is closed under multiplication and the adjoint operation. It is, however, not a group.*

(ii) Every permutation matrix is doubly stochastic and is an extreme point of the convex set of all doubly stochastic matrices. (Later we will prove Birkhoff's Theorem, which says that all extreme points of this convex set are permutation matrices.)

Exercise II.1.7 *Let A be a doubly stochastic matrix. Show that all eigenvalues of A have modulus less than or equal to 1, that 1 is an eigenvalue of A, and that $||A|| = 1$.*

Exercise II.1.8 *If A is doubly stochastic, then*

$$|Ax| \leq A(|x|),$$

where, as usual, $|x| = (|x_1|, \ldots, |x_n|)$ and we say that $x \leq y$ if $x_j \leq y_j$ for all j.

There is a close relationship between majorisation and doubly stochastic matrices. This is brought out in the next few theorems.

Theorem II.1.9 *A matrix A is doubly stochastic if and only if $Ax \prec x$ for all vectors x.*

Proof. Let $Ax \prec x$ for all x. First choosing x to be e and then $e_i = (0, 0, \ldots, 1, 0, \ldots, 0), 1 \leq i \leq n$, one can easily see that A is doubly stochastic.

Conversely, let A be doubly stochastic. Let $y = Ax$. To prove $y \prec x$ we may assume, without loss of generality, that the coordinates of both x and

y are in decreasing order. (See Remark II.1.2 and Exercise II.1.6.) Now note that for any $k, 1 \le k \le n$, we have

$$\sum_{j=1}^{k} y_j = \sum_{j=1}^{k}\sum_{i=1}^{n} a_{ij}x_i.$$

If we put $t_i = \sum_{j=1}^{k} a_{ij}$, then $0 \le t_i \le 1$ and $\sum_{i=1}^{n} t_i = k$. We have

$$\begin{aligned}
\sum_{j=1}^{k} y_j - \sum_{j=1}^{k} x_j &= \sum_{i=1}^{n} t_i x_i - \sum_{i=1}^{k} x_i \\
&= \sum_{i=1}^{n} t_i x_i - \sum_{i=1}^{k} x_i + (k - \sum_{i=1}^{n} t_i)x_k \\
&= \sum_{i=1}^{k} (t_i - 1)(x_i - x_k) + \sum_{i=k+1}^{n} t_i(x_i - x_k) \\
&\le 0.
\end{aligned}$$

Further, when $k = n$ we must have equality here simply because A is doubly stochastic. Thus, $y \prec x$. ■

Note that if $x, y \in \mathbb{R}^2$ and $x \prec y$ then

$$(x_1, x_2) = (ty_1 + (1-t)y_2, (1-t)y_1 + ty_2) \text{ for some } 0 \le t \le 1.$$

Note also that if $x, y \in \mathbb{R}^n$ and x is obtained by averaging any two coordinates of y in the above sense while keeping the rest of the coordinates fixed, then $x \prec y$. More precisely, call a linear map T on $\mathbb{R}^n$ a **T-transform** if there exists $0 \le t \le 1$ and indices j, k such that

$$Ty = (y_1, \ldots, y_{j-1}, ty_j + (1-t)y_k, y_{j+1}, \ldots, (1-t)y_j + ty_k, y_{k+1}, \ldots, y_n).$$

Then, $Ty \prec y$ for all y.

Theorem II.1.10 *For $x, y \in \mathbb{R}^n$, the following statements are equivalent:*

(i) $x \prec y$.

(ii) x is obtained from y by a finite number of T-transforms.

(iii) x is in the convex hull of all vectors obtained by permuting the coordinates of y.

(iv) $x = Ay$ for some doubly stochastic matrix A.

Proof. When $n = 2$, then (i) $\Rightarrow$ (ii). We will prove this for a general n by induction. Assume that we have this implication for dimensions up to $n-1$. Let $x, y \in \mathbb{R}^n$. Since $x^{\downarrow}$ and $y^{\downarrow}$ can be obtained from x and y by permutations and each permutation is a product of transpositions – which are surely T-transforms, we can assume without loss of generality that $x_1 \geq x_2 \geq \cdots \geq x_n$ and $y_1 \geq y_2 \geq \cdots \geq y_n$. Now, if $x \prec y$, then $y_n \leq x_1 \leq y_1$. Choose k such that $y_k \leq x_1 \leq y_{k-1}$. Then $x_1 = ty_1 + (1-t)y_k$ for some $0 \leq t \leq 1$. Let

$$T_1 z = (tz_1 + (1-t)z_k, z_2, \ldots, z_{k-1}, (1-t)z_1 + tz_k, z_{k+1}, \ldots, z_n)$$

for all $z \in \mathbb{R}^n$. Then note that the first coordinate of $T_1 y$ is x_1. Let

$$\begin{aligned} x' &= (x_2, \ldots, x_n) \\ y' &= (y_2, \ldots, y_{k-1}, (1-t)y_1 + ty_k, y_{k+1}, \ldots, y_n). \end{aligned}$$

We will show that $x' \prec y'$. Since $y_1 \geq \cdots \geq y_{k-1} \geq x_1 \geq x_2 \geq \cdots \geq x_n$, we have for $2 \leq m \leq k-1$

$$\sum_{j=2}^{m} x_j \leq \sum_{j=2}^{m} y_j.$$

For $k \leq m \leq n$

$$\begin{aligned} \sum_{j=2}^{m} y'_j &= \sum_{j=2}^{k-1} y_j + [(1-t)y_1 + ty_k] + \sum_{j=k+1}^{m} y_j \\ &= \sum_{j=1}^{m} y_j - ty_1 + (t-1)y_k \\ &= \sum_{j=1}^{m} y_j - x_1 \geq \sum_{j=1}^{m} x_j - x_1 = \sum_{j=2}^{m} x_j. \end{aligned}$$

The last inequality is an equality when $m = n$ since $x \prec y$. Thus $x' \prec y'$. So by the induction hypothesis there exist a finite number of T-transforms $T_2, \ldots, T_r$ on $\mathbb{R}^{n-1}$ such that $x' = (T_r \cdots T_2)y'$. We can regard each of them as a T-transform on $\mathbb{R}^n$ if we prohibit them from touching the first coordinate of any vector. We then have

$$(T_r \cdots T_1)y = (T_r \cdots T_2)(x_1, y') = (x_1, x') = x,$$

and that is what we wanted to prove.

Now note that a T-transform is a convex combination of the identity map and some permutation. So a product of such maps is a convex combination of permutations. Hence (ii) $\Rightarrow$ (iii). The implication (iii) $\Rightarrow$ (iv) is obvious, and (iv) $\Rightarrow$ (i) is a consequence of Theorem II.1.9. ∎

A consequence of the above theorem is that the set $\{x : x \prec y\}$ is the convex hull of all points obtained from y by permuting its coordinates.

Exercise II.1.11 *If $U = (u_{ij})$ is a unitary matrix, then the matrix $(|u_{ij}|^2)$ is doubly stochastic. Such a doubly stochastic matrix is called* **unitary-stochastic**; *it is called* **orthostochastic** *if U is real orthogonal. Show that if $x = Ay$ for some doubly stochastic matrix A, then there exists an orthostochastic matrix B such that $x = By$. (Use induction.)*

Exercise II.1.12 *Let A be an $n \times n$ Hermitian matrix. Let* diag (A) denote the vector whose coordinates are the diagonal entries of A and $\lambda(A)$ the vector whose coordinates are the eigenvalues of A specified in any order. Show that

$$\text{diag}\,(A) \prec \lambda(A). \tag{II.14}$$

This is sometimes referred to as **Schur's Theorem.**

Exercise II.1.13 *Use the majorisation (II.14) to prove that if $\lambda_j^{\downarrow}(A)$ denote the eigenvalues of an $n \times n$ Hermitian matrix arranged in decreasing order then for all $k = 1, 2, \ldots, n$*

$$\sum_{j=1}^{k} \lambda_j^{\downarrow}(A) = \max \sum_{j=1}^{k} \langle x_j, Ax_j \rangle, \tag{II.15}$$

where the maximum is taken over all orthonormal k-tuples of vectors $\{x_1, \ldots, x_k\}$ in $\mathbb{C}^n$. This is the **Ky Fan's maximum principle**. *(See Problem I.6.15 also.) Show that the majorisation (II.14) can be derived from (II.15). The two statements are, thus, equivalent.*

Exercise II.1.14 *Let A, B be Hermitian matrices. Then for all $k = 1, 2, \ldots, n$*

$$\sum_{j=1}^{k} \lambda_j^{\downarrow}(A+B) \le \sum_{j=1}^{k} \lambda_j^{\downarrow}(A) + \sum_{j=1}^{k} \lambda_j^{\downarrow}(B). \tag{II.16}$$

Exercise II.1.15 *For any matrix A, let $\tilde{A}$ be the Hermitian matrix*

$$\tilde{A} = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}. \tag{II.17}$$

Then the eigenvalues of $\tilde{A}$ are the singular values of A together with their negatives. Denote the singular values of A arranged in decreasing order by $s_1(A), \ldots, s_n(A)$. Show that for any two $n \times n$ matrices A, B and for any $k = 1, 2, \ldots, n$

$$\sum_{j=1}^{k} s_j(A+B) \le \sum_{j=1}^{k} s_j(A) + \sum_{j=1}^{k} s_j(B). \tag{II.18}$$

When $k = 1$, this is just the triangle inequality for the operator norm $||A||$. For each $1 \le k \le n$, define $||A||_{(k)} = \sum_{j=1}^{k} s_j(A)$. From (II.18) it follows that $||A||_{(k)}$ defines a norm. These norms are called the **Ky Fan k-norms.**

II.2 Birkhoff's Theorem

We start with a combinatorial problem known as the **Matching Problem.**

Let $B = \{b_1, \ldots, b_n\}$ and $G = \{g_1, \ldots, g_n\}$ be two sets of n elements each, and let R be a subset of $B \times G$. When does there exist a bijection f from B to G whose graph is contained in R? This is called the Matching Problem or the **Marriage Problem** for the following reason. Think of B as a set of boys, G as a set of girls, and $(b_i, g_j) \in R$ as saying that the boy b_i knows the girl g_j. Then the above question can be phrased as: when can one arrange a monogamous marriage in which each boy gets married to a girl he knows? We will call such a matching a **compatible matching**.

For each i let $G_i = \{g_j : (b_i, g_j) \in R\}$. This represents the set of girls whom the boy b_i knows. For each k-tuple of indices $1 \le i_1 < \cdots < i_k \le n$, let $G_{i_1 \cdots i_k} = \bigcup_{r=1}^{k} G_{i_r}$. This represents the set of girls each of whom are known to one of the boys $b_{i_1}, \ldots, b_{i_k}$. Clearly a necessary condition for a compatible matching to be possible is that $|G_{i_1 \cdots i_k}| \ge k$ for all $k = 1, 2, \ldots, n$. Hall's Marriage Theorem says that this condition is sufficient as well.

Theorem II.2.1 *(Hall) A compatible matching between B and G can be found if and only if*

$$|G_{i_1 \cdots i_k}| \ge k, \tag{II.19}$$

for all $1 \le i_1 < \cdots < i_k \le n, \; k = 1, 2, \ldots, n.$

Proof. Only the sufficiency of the condition needs to be proved. This is done by induction on n. Obviously, the Theorem is true when $n = 1$.

First assume that we have

$$|G_{i_1 \cdots i_k}| \ge k + 1,$$

for all $1 \le i_1 < \cdots < i_k \le n, 1 \le k < n$. In other words, if $1 \le k < n$, then every set of k boys together knows at least $k+1$ girls. Pick up any boy and marry him to one of the girls he knows. This leaves $n-1$ boys and $n-1$ girls; condition (II.19) still holds, and hence the remaining boys and girls can be compatibly matched.

If the above assumption is not met, then there exist k indices $i_1, \ldots, i_k$, $k < n$, for which

$$|G_{i_1 \cdots i_k}| = k.$$

In other words, there exist k boys who together know exactly k girls. By the induction hypothesis these k boys and girls can be compatibly matched. Now we are left with $n-k$ unmarried boys and as many unmarried girls. If some set of h of these boys knew less than h of these remaining girls, then together with the earlier k these $h+k$ boys would have known less than $h+k$ girls. (The earlier k boys did not know any of the present $n-k$ maidens.)

So, condition (II.19) is satisfied for the remaining $n-k$ boys and girls who can now be compatibly married by the induction hypothesis. ■

Exercise II.2.2 *(The König-Frobenius Theorem) Let $A = (a_{ij})$ be an $n \times n$ matrix. If σ is a permutation on n symbols, the set $\{a_{1\sigma(1)}, a_{2\sigma(2)}, \ldots, a_{n\sigma(n)}\}$ is called a* **diagonal** *of A . Each diagonal contains exactly one element from each row and from each column of A. Show that the following two statements are equivalent:*

(i) every diagonal of A contains a zero element.

(ii) A has a $k \times \ell$ submatrix with all entries zero for some k, ℓ such that $k + \ell > n$.

One can see that the statement of the König-Frobenius Theorem is *equivalent* to that of Hall's Theorem.

Theorem II.2.3 *(Birkhoff's Theorem) The set of $n \times n$ doubly stochastic matrices is a convex set whose extreme points are the permutation matrices.*

Proof. We have already made a note of the easy part of this theorem in Exercise II.1.6. The harder part is showing that every extreme point is a permutation matrix. For this we need to show that each doubly stochastic matrix is a convex combination of permutation matrices.

This is proved by induction on the number of positive entries of the matrix. Note that if A is doubly stochastic, then it has at least n positive entries. If the number of positive entries is exactly n, then A is a permutation matrix.

We first show that if A is doubly stochastic, then A has at least one diagonal with no zero entry. Choose any $k \times \ell$ submatrix of zeroes that A might have. We can find permutation matrices P_1, P_2 such that $P_1 A P_2$ has the form

$$P_1 A P_2 = \begin{bmatrix} O & B \\ C & D \end{bmatrix},$$

where O is a $k \times \ell$ matrix with all entries zero. Since $P_1 A P_2$ is again doubly stochastic, the rows of B and the columns of C each add up to 1. Hence $k + \ell \leq n$. So at least one diagonal of A must have all its entries positive, by the König-Frobenius Theorem.

Choose any such positive diagonal and let a be the smallest of the elements of this diagonal. If A is not a permutation matrix, then $a < 1$. Let P be the permutation matrix obtained by putting ones on this diagonal and let

$$B = \frac{A - aP}{1 - a}.$$

Then B is doubly stochastic and has at least one more zero entry than A has. So by the induction hypothesis B is a convex combination of permutation matrices. Hence so is A, since $A = (1 - a)B + aP$. ■

Remark. There are $n!$ permutation matrices of size n. Birkhoff's Theorem tells us that every $n \times n$ doubly stochastic matrix is a convex combination of these $n!$ matrices. This number can be reduced as a consequence of a general theorem of Carathéodory. This says that if X is a subset of an m-dimensional linear variety in $\mathbb{R}^N$, then any point in the convex hull of X can be expressed as a convex combination of at most $m+1$ points of X. Using this theorem one sees that every $n \times n$ doubly stochastic matrix can be expressed as a convex combination of at most $n^2 - 2n + 2$ permutation matrices.

Doubly substochastic matrices defined below are related to weak majorisation in the same way as doubly stochastic matrices are related to majorisation.

A matrix $B = (b_{ij})$ is called **doubly substochastic** if

$$\begin{aligned} b_{ij} &\geq 0 \quad &&\text{for all} \quad i,j, \\ \sum_{i=1}^{n} b_{ij} &\leq 1 \quad &&\text{for all} \quad j, \\ \sum_{j=1}^{n} b_{ij} &\leq 1 \quad &&\text{for all} \quad i. \end{aligned}$$

Exercise II.2.4 *B is doubly substochastic if it is positivity-preserving, $Be \leq e$, and $B^*e \leq e$.*

Exercise II.2.5 *Every square submatrix of a doubly stochastic matrix is doubly substochastic. Conversely, every doubly substochastic matrix B can be dilated to a doubly stochastic matrix A. Moreover, if B is an $n \times n$ matrix, then this dilation A can be chosen to have size at most $2n \times 2n$. Indeed, if R and C are the diagonal matrices whose jth diagonal entries are the sums of the jth rows and the jth columns of B, respectively, then*

$$A = \begin{pmatrix} B & I-R \\ I-C & B^* \end{pmatrix}$$

is a doubly stochastic matrix.

Exercise II.2.6 *The set of all $n \times n$ doubly substochastic matrices is convex; its extreme points are matrices having at most one entry 1 in each row and each column and all other entries zero.*

Exercise II.2.7 *A matrix B with nonnegative entries is doubly substochastic if and only if there exists a doubly stochastic matrix A such that $b_{ij} \leq a_{ij}$ for all $i, j = 1, 2, \ldots, n$.*

Our next theorem connects doubly substochastic matrices to weak majorisation.

Theorem II.2.8 *(i) Let x, y be two vectors with nonnegative coordinates. Then $x \prec_w y$ if and only if $x = By$ for some doubly substochastic matrix B.*

(ii) Let $x, y \in \mathbb{R}^n$. Then $x \prec_w y$ if and only if there exists a vector u such that $x \leq u$ and $u \prec y$.

Proof. If $x, u \in \mathbb{R}^n$ and $x \leq u$, then clearly $x \prec_w u$. So, if in addition $u \prec y$, then $x \prec_w y$.

Now suppose that x, y are nonnegative vectors and $x = By$ for some doubly substochastic matrix B. By Exercise II.2.7 we can find a doubly stochastic matrix A such that $b_{ij} \leq a_{ij}$ for all i, j. Then $x = By \leq Ay$. Hence, $x \prec_w y$.

Conversely, let x, y be nonnegative vectors such that $x \prec_w y$. We want to prove that there exists a doubly substochastic matrix B for which $x = By$. If $x = 0$, we can choose $B = 0$, and if $x \prec y$, we can even choose B to be doubly stochastic by Theorem II.1.10. So, assume that neither of these is the case. Let r be the smallest of the positive coordinates of x, and let $s = \Sigma\, y_j - \Sigma\, x_j$. By assumption $s > 0$. Choose a positive integer m such that $r \geq s/m$. Dilate both vectors x and y to $(n+m)$-dimensional vectors x', y' defined as

$$
\begin{aligned}
x' &= (x_1, \ldots, x_n, s/m, \ldots, s/m), \\
y' &= (y_1, \ldots, y_n, 0, \ldots, 0).
\end{aligned}
$$

Then $x' \prec y'$. Hence $x' = Ay'$ for some doubly stochastic matrix of size $n+m$. Let B be the $n \times n$ submatrix of A sitting in the top left corner. Then B is doubly substochastic and $x = By$. This proves (i).

Finally, let $x, y \in \mathbb{R}^n$ and $x \prec_w y$. Choose a positive number t so that $x + te$ and $y + te$ are both nonnegative, where $e = (1, 1, \ldots, 1)$. We still have $x + te \prec_w y + te$. So, by (i) there exists a doubly substochastic matrix B such that $x + te = B(y + te)$. By Exercise II.2.7 we can find a doubly stochastic matrix A such that $b_{ij} \leq a_{ij}$ for all i, j. But then $x + te \leq A(y + te) = Ay + te$. Hence, if $u = Ay$, then $x \leq u$ and $u \prec y$. ∎

Exercise II.2.9 *A matrix A is doubly substochastic if and only if for every $x \geq 0$ we have $Ax \geq 0$ and $Ax \prec_w x$. (Compare with Theorem II.1.9.)*

Exercise II.2.10 *Let $x, y \in \mathbb{R}^n$ and let $x \geq 0, y \geq 0$. Then $x \prec_w y$ if and only if x is in the convex hull of the $2^n n!$ points obtained from y by permutations and sign changes of its coordinates (i.e., vectors of the form $(\pm y_{\sigma(1)}, \pm y_{\sigma(2)}, \ldots, \pm y_{\sigma(n)})$, where σ is a permutation).*

II.3 Convex and Monotone Functions

In this section we will study maps from $\mathbb{R}^n$ to $\mathbb{R}^m$ that preserve various orders.

Let $f : \mathbb{R} \to \mathbb{R}$ be any function. We will denote the map induced by f on $\mathbb{R}^n$ also by f; i.e., $f(x) = (f(x_1), \ldots, f(x_n))$ for $x \in \mathbb{R}^n$. An elementary and useful characterisation of majorisation is the following.

Theorem II.3.1 *Let $x, y \in \mathbb{R}^n$. Then the following two conditions are equivalent:*

(i) $x \prec y$.

(ii) $\operatorname{tr} \varphi(x) \leq \operatorname{tr} \varphi(\mathrm{y})$ *for all convex functions φ from $\mathbb{R}$ to $\mathbb{R}$.*

Proof. Let $x \prec y$. Then $x = Ay$ for some doubly stochastic matrix A. So $x_i = \sum_{j=1}^{n} a_{ij} y_j$, where $a_{ij} \geq 0$ and $\sum_{j=1}^{n} a_{ij} = 1$. Hence for every convex function $\varphi, \varphi(x_i) \leq \sum_{j=1}^{n} a_{ij}\varphi(y_j)$. Hence $\sum_{i=1}^{n}\varphi(x_i) \leq \sum_{i,j} a_{ij}\varphi(y_j) = \sum_{j=1}^{n}\varphi(y_j)$. To prove the converse note that for each t the function $\varphi_t(x) = |x - t|$ is convex. Now apply Theorem II.1.3 (iii). ■

Exercise II.3.2 *For $x, y \in \mathbb{R}^n$ the following two conditions are equivalent:*

(i) $x \prec_w y$.

(ii) $\operatorname{tr} \varphi(x) \leq \operatorname{tr} \varphi(y)$ *for all monotonically increasing convex functions φ from $\mathbb{R}$ to $\mathbb{R}$.*

Note that in the two statements above it suffices to consider only continuous functions.

A real valued function φ on $\mathbb{R}^n$ is called **Schur-convex** or **S-convex** if

$$x \prec y \quad \Rightarrow \quad \varphi(x) \leq \varphi(y). \tag{II.20}$$

(This terminology might seem somewhat inappropriate because the condition (II.20) expresses preservation of order rather than convexity. However, the above two propositions do show that ordinary convex functions are related to this notion. Also, if $x \prec y$, then x is obtained from y by an averaging procedure. The condition (II.20) says that the value of φ is diminished when such a procedure is applied to its argument. Later on, we will come across other notions of averaging, and corresponding notions of convexity.)

We will study more general maps that include Schur-convex maps.

Consider maps $\Phi : \mathbb{R}^n \to \mathbb{R}^m$. The domain of Φ will be either all of $\mathbb{R}^n$ or some convex set invariant under coordinate permutations of its elements. Such a map will be called **monotone increasing** if

$$x \leq y \quad \Rightarrow \quad \Phi(x) \leq \Phi(y),$$

monotone decreasing if

$$-\Phi \text{ is monotone increasing,}$$

convex if

$$\Phi(tx + (1-t)y) \leq t\Phi(x) + (1-t)\Phi(y), \quad 0 \leq t \leq 1,$$

concave if

$$-\Phi \quad \text{is convex,}$$

isotone if

$$x \prec y \quad \Longrightarrow \quad \Phi(x) \prec_w \Phi(y),$$

strongly isotone if

$$x \prec_w y \quad \Longrightarrow \quad \Phi(x) \prec_w \Phi(y),$$

and **strictly isotone** if

$$x \prec y \quad \Longrightarrow \quad \Phi(x) \prec \Phi(y).$$

Note that when $m = 1$ isotone maps are precisely the Schur-convex maps.

The next few propositions provide examples of such maps. We will denote by S_n the group of $n \times n$ permutation matrices.

Theorem II.3.3 *Let $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ be a convex map. Suppose that for any $P \in S_n$ there exists $P' \in S_m$ such that*

$$\Phi(Px) = P'\Phi(x) \quad \text{for all} \quad x \in \mathbb{R}^n. \tag{II.21}$$

Then Φ is isotone. In addition, if Φ is monotone increasing, then Φ is strongly isotone.

Proof. Let $x \prec y$ in $\mathbb{R}^n$. By Theorem II.1.10 there exist $P_1, \ldots, P_N$ in S_n and positive real numbers $t_1, \ldots, t_N$ with $\Sigma t_j = 1$ such that

$$x = \Sigma t_j P_j y.$$

So, by the convexity of Φ and the property (II.21)

$$\Phi(x) \leq \Sigma t_j \Phi(P_j y) = \Sigma t_j P_j' \Phi(y) = z, \text{say.}$$

Then $z \prec \Phi(y)$ and $\Phi(x) \leq z$. So $\Phi(x) \prec_w \Phi(y)$. This proves that Φ is isotone.

Suppose Φ is also monotone increasing. Let $u \prec_w y$. Then by Theorem II.2.8 there exists x such that $u \leq x \prec y$. Hence $\Phi(u) \leq \Phi(x)$ and $\Phi(x) \prec_w \Phi(y)$. So, $\Phi(u) \prec_w \Phi(y)$. This proves Φ is strongly isotone. ∎

Corollary II.3.4 *If $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, then the induced map $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ is isotone. If φ is convex and monotone on $\mathbb{R}$, then the induced map is strongly isotone on $\mathbb{R}^n$.*

Note that one part of Theorem II.3.1 and Exercise II.3.2 is subsumed by the above corollary.

Example II.3.5 *From the above results we can conclude that*

(i) $x \prec y$ *in* $\mathbb{R}^n \Rightarrow |x| \prec_w |y|$.

(ii) $x \prec y$ *in* $\mathbb{R}^n \Rightarrow x^2 \prec_w y^2$.

(iii) $x \prec_w y$ *in* $\mathbb{R}^n_+ \Rightarrow x^p \prec_w y^p$ *for* $p > 1$.

(iv) $x \prec_w y$ *in* $\mathbb{R}^n \Rightarrow x^+ \prec_w y^+$.

(v) If φ is any function such that $\varphi(e^t)$ is convex and monotone increasing in t, then $\log x \prec_w \log y$ *in* $\mathbb{R}^n_+ \Rightarrow \varphi(x) \prec_w \varphi(y)$.

(vi) $\log x \prec_w \log y$ *in* $\mathbb{R}^n_+ \Rightarrow x \prec_w y$.

(vii) For $x, y \in \mathbb{R}^n_+$

$$\prod_{j=1}^{k} x_j^{\downarrow} \leq \prod_{j=1}^{k} y_j^{\downarrow}, 1 \leq k \leq n, \Rightarrow \sum_{j=1}^{k} x_j^{\downarrow} \leq \sum_{j=1}^{k} y_j^{\downarrow}, \; 1 \leq k \leq n.$$

Here $\mathbb{R}^n_+$ stands for the collection of vectors $x \geq 0$ (or, at places, $x > 0$). All functions are understood in the coordinatewise sense. Thus, e.g., $|x| = (|x_1|, \ldots, |x_n|)$.

As an application we have the following very useful theorem.

Theorem II.3.6 *(Weyl's Majorant Theorem) Let A be an $n \times n$ matrix with singular values $s_1 \geq \cdots \geq s_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$ arranged in such a way that $|\lambda_1| \geq \cdots \geq |\lambda_n|$. Then for every function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$, such that $\varphi(e^t)$ is convex and monotone increasing in t, we have*

$$(\varphi(|\lambda_1|), \ldots, \varphi(|\lambda_n|)) \prec_w (\varphi(s_1), \ldots, \varphi(s_n)). \tag{II.22}$$

In particular, we have

$$(|\lambda_1|^p, \ldots, |\lambda_n|^p) \prec_w (s_1^p, \ldots, s_n^p), \tag{II.23}$$

for all $p \geq 0$.

Proof. The spectral radius of a matrix is bounded by its operator norm. Hence,

$$|\lambda_1| \leq ||A|| = s_1.$$

Apply this argument to the antisymmetric tensor powers $\wedge^k A$. This gives

$$\prod_{j=1}^{k} |\lambda_j| \leq \prod_{j=1}^{k} s_j, \quad 1 \leq k \leq n. \tag{II.24}$$

Now use the assertion of II.3.5 (vii). ■

Note that we have

$$\prod_{j=1}^{n} |\lambda_j| = \prod_{j=1}^{n} s_j, \tag{II.25}$$

both the expressions being equal to $(\det A^*A)^{1/2}$.

Remark II.3.7 *Returning to Theorem II.3.3, we note that when $m = 1$ the condition (II.21) just says that Φ is* **permutation invariant**; *i.e.,*

$$\Phi(Px) = \Phi(x), \tag{II.26}$$

for all $x \in \mathbb{R}^n$ and $P \in S_n$. So, in this case Theorem II.3.3 says that if a function $\Phi : \mathbb{R}^n \to \mathbb{R}$ is convex and permutation invariant, then it is isotone (i.e., Schur-convex).

Also note that every isotone function Φ from $\mathbb{R}^n$ to $\mathbb{R}$ has to be permutation invariant because Px and x majorise each other and hence isotony of Φ implies equality of $\Phi(Px)$ and $\Phi(x)$ in this case.

However, we will see that not every isotone function from $\mathbb{R}^n$ to $\mathbb{R}$ (i.e. not every Schur-convex function) is convex.

Exercise II.3.8 *Let $\Psi : \mathbb{R}^n \to \mathbb{R}$ be any convex function and let $\Phi(x) = \max_{P \in S_n} \Psi(Px)$. Prove that Φ is isotone. If, in addition, Ψ is monotone increasing, then Φ is strongly isotone.*

Exercise II.3.9 *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex. For each $k = 1, 2, \ldots, n$, define functions $\varphi^{(k)} : \mathbb{R}^n \to \mathbb{R}$ by*

$$\varphi^{(k)}(x) = \max_{\sigma} \sum_{j=1}^{k} \varphi(x_{\sigma(j)}),$$

where σ runs over all permutations on n symbols. Then $\varphi^{(k)}$ is isotone. If, in addition, φ is monotone increasing, then $\varphi^{(k)}$ is strongly isotone. Note that this applies, in particular, to

$$\varphi^{(n)}(x) = \sum_{j=1}^{n} \varphi(x_j) = \text{tr}\ \varphi(\text{x}).$$

Compare this with Theorem II.3.1. The special choice $\varphi(t) = t$ gives $\varphi^{(k)}(x) = \sum_{j=1}^{k} x_j^{\downarrow}$.

Example II.3.10 *For $x \in \mathbb{R}^n$ let $\bar{x} = \frac{1}{n}\Sigma x_j$. Let*

$$V(x) = \frac{1}{n}\sum_j (x_j - \bar{x})^2.$$

This is called the **variance** *function. Since the maps $x_j \to (x_j - \bar{x})^2$ are convex, $V(x)$ is isotone (i.e., Schur-convex).*

Example II.3.11 *For $x \in \mathbb{R}^n_+$ let*

$$H(x) = -\sum_j x_j \log x_j,$$

where by convention we put $t \log t = 0$, if $t = 0$. Then H is called the **entropy** *function. Since the function $f(t) = t \log t$ is convex for $t \geq 0$, we see that $-H(x)$ is isotone. (This is sometimes expressed by saying that the entropy function is anti-isotone or Schur-concave on $\mathbb{R}^n_+$.) In particular, if $x_j \geq 0$ and $\Sigma x_j = 1$ we have*

$$H(1, 0, \ldots, 0) \leq H(x_1, \ldots, x_n) \leq H(\frac{1}{n}, \ldots, \frac{1}{n}),$$

which is a basic fact about entropy.

Example II.3.12 *For $p \geq 1$ the function*

$$\Phi(x) = \sum_{j=1}^{n} (x_j + \frac{1}{x_j})^p$$

is isotone on $\mathbb{R}^n_+$. In particular, if $x_j > 0$ and $\Sigma x_j = 1$, we have

$$\frac{(n^2+1)^p}{n^{p-1}} \leq \sum_{j=1}^{n} (x_j + \frac{1}{x_j})^p.$$

Example II.3.13 *A function $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ is called a* **symmetric gauge function** *if*

(i) Φ is a norm on the real vector space $\mathbb{R}^n$,

(ii) $\Phi(Px) = \Phi(x)$ for all $x \in \mathbb{R}^n, P \in S_n$,

(iii) $\Phi(\varepsilon_1 x_1, \ldots, \varepsilon_n x_n) = \Phi(x_1, \ldots, x_n)$ if $\varepsilon_j = \pm 1$,

(iv) $\Phi(1, 0, \ldots, 0) = 1$.

(The last condition is an inessential normalisation.) Examples of symmetric gauge functions are

$$\begin{aligned} \Phi_p(x) &= (\sum_{j=1}^{n} |x_j|^p)^{1/p}, \quad 1 \leq p < \infty, \\ \Phi_\infty(x) &= \max_{1 \leq j \leq n} |x_j|. \end{aligned}$$

These norms are commonly used in functional analysis. If the coordinates of x are arranged so as to have $|x_1| \geq \cdots \geq |x_n|$, then

$$\Phi_{(k)}(x) = \sum_{j=1}^{k} |x_j|, \qquad 1 \leq k \leq n$$

is also a symmetric gauge function. This is a consequence of the majorisations (II.29) and (i) in Examples II.3.5.

Every symmetric gauge function is convex on $\mathbb{R}^n$ and is monotone on $\mathbb{R}^n_+$ (Problem II.5.11). Hence by Theorem II.3.3 it is strongly isotone; i.e.,

$$x \prec_w y \text{ in } \mathbb{R}^n_+ \Rightarrow \Phi(x) \leq \Phi(y).$$

For differentiable functions there are necessary and sufficient conditions characterising Schur-convexity:

Theorem II.3.14 *A differentiable function $\Phi : \mathbb{R}^n \to \mathbb{R}$ is isotone if and only if*

(i) Φ is permutation invariant, and

(ii) for each $x \in \mathbb{R}^n$ and for all i, j

$$(x_i - x_j)\left(\frac{\partial \Phi}{\partial x_i}(x) - \frac{\partial \Phi}{\partial x_j}(x)\right) \geq 0.$$

Proof. We have already observed that every isotone function is permutation invariant. To see that it also satisfies (ii), let $i = 1, j = 2$, without any loss of generality. For $0 \leq t \leq 1$ let

$$x(t) = ((1-t)x_1 + tx_2, tx_1 + (1-t)x_2, x_3, \ldots, x_n). \tag{II.27}$$

Then $x(t) \prec x = x(0)$. Hence $\Phi(x(t)) \leq \Phi(x(0))$, and therefore

$$0 \geq \left[\frac{d}{dt}\Phi(x(t))\right]_{t=0} = -(x_1 - x_2)\left(\frac{\partial \Phi}{\partial x_1}(x) - \frac{\partial \Phi}{\partial x_2}(x)\right).$$

This proves (ii).

Conversely, suppose Φ satisfies (i) and (ii). We want to prove that $\Phi(u) \leq \Phi(x)$ if $u \prec x$. By Theorem II.1.10 and the permutation invariance of Φ we may assume that

$$u = ((1-s)x_1 + sx_2, sx_1 + (1-s)x_2, x_3, \ldots, x_n)$$

for some $0 \leq s \leq \frac{1}{2}$. Let $x(t)$ be as in (II.27). Then

$$\Phi(u) - \Phi(x) \quad = \quad \int_0^s \frac{d}{dt}\Phi(x(t))dt$$

$$\begin{aligned} &= -\int_0^s (x_1 - x_2)\left[\frac{\partial\Phi}{\partial x_1}(x(t)) - \frac{\partial\Phi}{\partial x_2}(x(t))\right] dt \\ &= -\int_0^s \frac{x(t)_1 - x(t)_2}{1-2t}\left[\frac{\partial\Phi}{\partial x_1}(x(t)) - \frac{\partial\Phi}{\partial x_2}(x(t))\right] dt \\ &\le 0, \end{aligned}$$

because of (ii) and the condition $0 \le s \le \frac{1}{2}$. ■

Example II.3.15 *(A Schur-convex function that is not convex) Let $\Phi : I^2 \to \mathbb{R}$, where $I = (0,1)$, be the function*

$$\Phi(x_1, x_2) = \log(\frac{1}{x_1} - 1) + \log(\frac{1}{x_2} - 1).$$

Using Theorem II.3.14 one can check that Φ is Schur-convex on the set

$$\{x : x \in I^2, x_1 + x_2 \le 1\}.$$

However, the function $\log(\frac{1}{t} - 1)$ is convex on $(0, \frac{1}{2}]$ but not on $[\frac{1}{2}, 1)$.

Example II.3.16 *(The elementary symmetric polynomials) For each $k = 1, 2, \cdots, n$, let $S_k : \mathbb{R}^n \to \mathbb{R}$ be the functions*

$$S_k(x) = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} x_{i_1} x_{i_2} \cdots x_{i_k}.$$

These are called the elementary symmetric polynomials of the n variables $x_1, \ldots, x_n$. These are invariant under permutations. We have the identities

$$\frac{\partial}{\partial x_j} S_k(x_1, \ldots, x_n) = S_{k-1}(x_1, \ldots, \hat{x}_j, \ldots, x_n)$$

and

$$\begin{aligned} &S_k(x_1, \ldots, \hat{x}_i, \ldots, x_n) - S_k(x_1, \ldots, \hat{x}_j, \ldots, x_n) \\ &\quad = (x_j - x_i) S_{k-1}(x_1, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, x_n), \end{aligned}$$

where the circumflex indicates that the term below it has been omitted. Using these one finds via Theorem II.3.14 that each S_k is Schur-concave; i.e., $-S_k$ is isotone, on $\mathbb{R}^n_+$.

The special case $k = n$ says that if $x, y \in \mathbb{R}^n_+$ and $x \prec y$, then $\prod_{j=1}^n x_j \ge \prod_{j=1}^n y_j$.

Theorem II.3.17 *(The Hadamard Determinant Theorem) If A is an $n \times n$ positive matrix, then*

$$\det A \le \prod_{j=1}^n a_{jj}.$$

Proof. Use Schur's Theorem (Exercise II.1.12) and the above statement about the Schur-concavity of the function $f(x) = \prod_j x_j$ on $\mathbb{R}^n_+$. ■

More generally, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of a positive matrix A, we have for $k = 1, 2, \ldots, n$

$$S_k(\lambda_1, \ldots, \lambda_n) \leq S_k(a_{11}, \ldots, a_{nn}). \tag{II.28}$$

Exercise II.3.18 *If A is an $m \times n$ complex matrix, then*

$$\det(AA^*) \leq \prod_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2.$$

(See Exercise I.1.3.)

Exercise II.3.19 *Show that the ratio $S_k(x)/S_{k-1}(x)$ is Schur-concave on the set of positive vectors for $k = 2, \ldots, n$. Hence, if A is a positive matrix, then*

$$\begin{aligned} \frac{S_n(a_{11}, \ldots, a_{nn})}{S_n(\lambda_1, \ldots, \lambda_n)} &\geq \frac{S_{n-1}(a_{11}, \ldots, a_{nn})}{S_{n-1}(\lambda_1, \ldots, \lambda_n)} \geq \cdots \geq \frac{S_1(a_{11}, \ldots, a_{nn})}{S_1(\lambda_1, \ldots, \lambda_n)} \\ &= \frac{\operatorname{tr} A}{\operatorname{tr} A} = 1. \end{aligned}$$

Proposition II.3.20 *If A is an $n \times n$ positive definite matrix, then*

$$(\det A)^{1/n} = \min \left\{ \frac{\operatorname{tr} AB}{n} : B \textit{ is positive and } \det B = 1 \right\}.$$

If A is positive semidefinite, then the same relation holds with min *replaced by* inf.

Proof. It suffices to prove the statement about positive definite matrices; the semidefinite case follows by a continuity argument. Using the spectral theorem and the cyclicity of the trace, the general case of the proposition can be reduced to the special case when A is diagonal. So, let A be diagonal with diagonal entries $\lambda_1, \ldots, \lambda_n$. Then, using the arithmetic-geometric mean inequality and Theorem II.3.17 we have

$$\frac{\operatorname{tr} AB}{n} = \frac{1}{n} \sum_j \lambda_j b_{jj} \geq (\prod_j \lambda_j)^{1/n} (\prod_j b_{jj})^{1/n} \geq (\det\ A)^{1/n} (\det\ B)^{1/n},$$

for every positive matrix B. Hence, $\frac{\operatorname{tr} AB}{n} \geq (\det\ A)^{1/n}$ if $\det\ B = 1$. When $B = (\det\ A)^{1/n} A^{-1}$ this becomes an equality. ■

Corollary II.3.21 *(The Minkowski Determinant Theorem) If A, B are $n \times n$ positive matrices then*

$$(\det(A + B))^{1/n} \geq (\det\ A)^{1/n} + (\det\ B)^{1/n}.$$

II.4 Binary Algebraic Operations and Majorisation

For $x \in \mathbb{R}^n$ we have seen in Section II.1 that

$$\sum_{j=1}^{k} x_j^{\downarrow} = \max_{|I|=k} \langle x, e_I \rangle.$$

It follows that if $x, y \in \mathbb{R}^n$, then

$$x + y \prec x^{\downarrow} + y^{\downarrow}. \tag{II.29}$$

In this section we will study majorisation relations of this form for sums, products, and other functions of two vectors.

A map $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is called **lattice superadditive** if

$$\varphi(s_1, t_1) + \varphi(s_2, t_2) \le \varphi(s_1 \vee s_2, t_1 \vee t_2) + \varphi(s_1 \wedge s_2, t_1 \wedge t_2). \tag{II.30}$$

We will call a map φ **monotone** if it is either monotonically increasing or monotonically decreasing in each of its arguments.

In this section we will adopt the following notation. Given $\varphi : \mathbb{R}^2 \to \mathbb{R}$, we will denote by Φ the map from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$ defined as

$$\Phi(x, y) = (\varphi(x_1, y_1), \ldots, \varphi(x_n, y_n)). \tag{II.31}$$

Example II.4.1 *(i) $\varphi(s,t) = s + t$ is a monotone and lattice superadditive function on $\mathbb{R}^2$.*

(ii) $\varphi(s,t) = st$ is a monotone and lattice superadditive function on $\mathbb{R}^2_+$.

For (i) above we have

$$\Phi(x, y) = (x_1 + y_1, \ldots, x_n + y_n) \quad \textit{for} \quad x, y \in \mathbb{R}^n,$$

and for (ii) we have

$$\Phi(x, y) = (x_1 y_1, \ldots, x_n y_n) \quad \textit{for} \quad x, y \in \mathbb{R}^n.$$

Theorem II.4.2 *If φ is monotone and lattice superadditive, then*

$$\Phi(x^{\downarrow}, y^{\uparrow}) \prec_w \Phi(x, y) \prec_w \Phi(x^{\downarrow}, y^{\downarrow}), \tag{II.32}$$

for all $x, y \in \mathbb{R}^n$.

Proof. Note that if we apply a coordinate permutation simultaneously to x and y, then $\Phi(x, y)$ undergoes the same coordinate permutation. The two outer terms in (II.32) remain unaffected and so do the majorisations. Hence, to prove (II.32) we may assume that $x = x^{\downarrow}$; i.e., $x_1 \ge x_2 \ge \cdots \ge x_n$. Next

note that we can find a finite sequence of vectors $u^{(0)}, u^{(1)}, \ldots, u^{(N)}$ such that

$$y^{\downarrow} = u^{(0)}, y^{\uparrow} = u^{(N)}, y = u^{(j)} \text{ for some } 1 \leq j \leq N,$$

and each $u^{(k+1)}$ is obtained from $u^{(k)}$ by interchanging two components in such a way as to move from the arrangement $y^{\downarrow}$ to $y^{\uparrow}$; i.e., we pick up two indices i, j such that

$$i < j \quad \text{and} \quad u_i^{(k)} > u_j^{(k)}$$

and interchange these two components to obtain the vector $u^{(k+1)}$. So, to prove (II.32) it suffices to prove

$$\Phi(x, u^{(k+1)}) \prec_w \Phi(x, u^{(k)}) \tag{II.33}$$

for $k = 0, 1, \ldots, N-1$. Since we have already assumed $x_1 \geq x_2 \geq \cdots \geq x_n$, to prove (II.33) we need to prove the two-dimensional majorisation

$$(\varphi(s_1, t_2), \varphi(s_2, t_1)) \prec_w (\varphi(s_1, t_1), \varphi(s_2, t_2)) \tag{II.34}$$

if $s_1 \geq s_2$ and $t_1 \geq t_2$. Now, by the definition of weak majorisation, this is equivalent to the two inequalities

$$\begin{aligned} \varphi(s_1, t_2) \vee \varphi(s_2, t_1) &\leq \varphi(s_1, t_1) \vee \varphi(s_2, t_2), \\ \varphi(s_1, t_2) + \varphi(s_2, t_1) &\leq \varphi(s_1, t_1) + \varphi(s_2, t_2), \end{aligned}$$

for $s_1 \geq s_2$ and $t_1 \geq t_2$. The first of these follows from the monotony of φ and the second from the lattice superadditivity. ■

Corollary II.4.3 *For $x, y \in \mathbb{R}^n$*

$$x^{\downarrow} + y^{\uparrow} \prec x + y \prec x^{\downarrow} + y^{\downarrow}. \tag{II.35}$$

For $x, y \in \mathbb{R}^n_+$

$$x^{\downarrow} \cdot y^{\uparrow} \prec_w x \cdot y \prec_w x^{\downarrow} \cdot y^{\downarrow}, \tag{II.36}$$

where $x \cdot y = (x_1 y_1, \ldots, x_n y_n)$.

Corollary II.4.4 *For $x, y \in \mathbb{R}^n$*

$$\langle x^{\downarrow}, y^{\uparrow} \rangle \leq \langle x, y \rangle \leq \langle x^{\downarrow}, y^{\downarrow} \rangle. \tag{II.37}$$

Proof. If $x \geq 0$ and $y \geq 0$, this follows from (II.36). In the general case, choose t large enough so that $x + te \geq 0$ and $y + te \geq 0$ and apply the special result. ■

The inequality (II.37) has a "mechanical" interpretation when $x \geq 0$ and $y \geq 0$. On a rod fixed at the origin, hang weights y_i at the points at distances x_i from the origin. The inequality (II.37) then says that the maximum moment is obtained if the heaviest weights are the farthest from the origin.

Exercise II.4.5 *The function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ defined as $\varphi(s,t) = s \wedge t$ is monotone and lattice superadditive on $\mathbb{R}^2$. Hence, for $x, y \in \mathbb{R}^n$*

$$x^{\downarrow} \wedge y^{\uparrow} \prec_w x \wedge y \prec_w x^{\downarrow} \wedge y^{\downarrow}.$$

II.5 Problems

Problem II.5.1. If a doubly stochastic matrix A is invertible and A^{-1} is also doubly stochastic, then A is a permutation.

Problem II.5.2. Let $y \in \mathbb{R}^n_+$. The set $\{x : x \in \mathbb{R}^n_+, x \prec_w y\}$ is the convex hull of the points $(r_1 y_{\sigma(1)}, \ldots, r_n y_{\sigma(n)})$, where σ varies over permutations and each r_j is either 0 or 1.

Problem II.5.3. Let $y \in \mathbb{R}^n$. The set $\{x \in \mathbb{R}^n : |x| \prec_w |y|\}$ is the convex hull of points of the form $(\varepsilon_1 y_{\sigma(1)}, \ldots, \varepsilon_n y_{\sigma(n)})$, where σ varies over permutations and each $\varepsilon_j = \pm 1$.

Problem II.5.4. Let $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ be a 2×2 block matrix and let $\mathcal{C}(A) = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22}. \end{pmatrix}$ If $U = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$, then we can write

$$\mathcal{C}(A) = \frac{1}{2}(A + UAU^*).$$

Let $\lambda(A)$ and $s(A)$ denote the n-vectors whose coordinates are the eigenvalues and the singular values of A, respectively.

Use (II.18) to show that

$$s(\mathcal{C}(A)) \prec_w s(A).$$

If A is Hermitian, use (II.16) to show that

$$\lambda(\mathcal{C}(A)) \prec \lambda(A).$$

Problem II.5.5. More generally, let $P_1, \ldots, P_r$ be a family of mutually orthogonal projections in $\mathbb{C}^n$ such that $\oplus P_j = I$. Then the operation of taking A to $\mathcal{C}(A) = \Sigma P_j A P_j$ is called a **pinching** of A. In an appropriate choice of basis this means that

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ A_{r1} & A_{r2} & \cdots & A_{rr} \end{bmatrix}, \quad \mathcal{C}(A) = \begin{bmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{rr} \end{bmatrix}.$$

Each such pinching is a product of $r-1$ pinchings of the 2×2 type introduced in Problem II.5.4. Show that for every pinching $\mathcal{C}$

$$s(\mathcal{C}(A)) \prec_w s(A) \tag{II.38}$$

for all matrices A, and

$$\lambda(\mathcal{C}(A)) \prec \lambda(A) \tag{II.39}$$

for all Hermitian matrices A. When $P_1, \ldots, P_n$ are the projections onto the coordinate axes, we get as a special case of (II.38) above

$$|\text{tr } A| \leq \sum_{j=1}^{n} s_j(A) = ||A||_1. \tag{II.40}$$

From (II.39) we get as a special case Schur's Theorem

$$\text{diag } (A) \prec \lambda(A),$$

which we saw before in Exercise II.1.12.

Problem II.5.6. Let A be positive. Then

$$\det A \leq \det \mathcal{C}(A), \tag{II.41}$$

for every pinching $\mathcal{C}$. This is called **Fischer's inequality** and includes the Hadamard Determinant Theorem as a special case.

Problem II.5.7. For each $k = 1, 2, \ldots, n$ and for each pinching $\mathcal{C}$ show that for positive definite A

$$S_k(\lambda(A)) \leq S_k(\lambda(\mathcal{C}(A))), \tag{II.42}$$

where $S_k(\lambda(A))$ denotes the kth elementary symmetric polynomial of the eigenvalues of A. This inequality, due to Ostrowski, includes (II.28) as a special case. It also includes (II.41) as a special case.

Problem II.5.8. If $\wedge^k A$ denotes the kth antisymmetric tensor power of A, then the above inequality can be written as

$$\text{tr } \wedge^k A \leq \text{tr } \wedge^k (\mathcal{C}(A)). \tag{II.43}$$

The operator inequality

$$\wedge^k A \leq \wedge^k (\mathcal{C}(A))$$

is **not** always true. This is shown by the following example. Let

$$A = \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and let $\mathcal{C}$ be the pinching induced by the pair of projections P and $I - P$. (The space $\wedge^2 \mathbb{C}^4$ is 6-dimensional.)

Problem II.5.9. Let $\{\lambda_1, \ldots, \lambda_n\}$, $\{\mu_1, \ldots, \mu_n\}$ be two n-tuples of complex numbers. Let

$$d(\lambda, \mu) = \min_{\sigma} \max_{1 \le j \le n} |\lambda_j - \mu_{\sigma(j)}|,$$

where the minimum is taken over all permutations on n symbols. This is called the **optimal matching distance** between the unordered n-tuples λ and μ. It defines a metric on the space $\mathbb{C}^n_{sym}$ of such n-tuples. Show that we also have

$$d(\lambda, \mu) = \max_{\substack{I, J \subset \{1,2,\ldots,n\} \\ |I|+|J|=n+1}} \min_{\substack{i \in I \\ j \in J}} |\lambda_i - \mu_j|.$$

Problem II.5.10. This problem gives a refinement of Hall's Theorem under an additional assumption that is often fulfilled in matching problems. In the notations introduced at the beginning of Section II.2, define

$$B_i = \{b_j : (b_j, g_i) \in R\}, \qquad 1 \le i \le n.$$

This is the set of boys known to the girl g_i. Let

$$B_{i_1 \cdots i_k} = \bigcup_{r=1}^{k} B_{i_r}, \quad 1 \le i_1 < \cdots < i_k \le n.$$

Suppose that for each $k = 1, 2, \ldots, \left[\frac{n}{2}\right]$ and for every choice of indices $1 \le i_1 < \cdots < i_k \le n$,

$$|G_{i_1 \cdots i_k}| \ge k \text{ and } |B_{i_1 \cdots i_k}| \ge k.$$

Show that then

$$|G_{i_1 \cdots i_k}| \ge k \text{ for all } k = 1, 2, \ldots, n, 1 \le i_1 < \cdots < i_k \le n.$$

Hence a compatible matching between B and G exists.

Problem II.5.11. (i) Show that every symmetric gauge function is continuous.

(ii) Show that if Φ is a symmetric gauge function, then $\Phi_\infty(x) \le \Phi(x) \le \Phi_1(x)$ for all $x \in \mathbb{R}^n$.

(iii) If Φ is a symmetric gauge function and $0 \le t_j \le 1$, then

$$\Phi(t_1 x_1, \ldots, t_n x_n) \le \Phi(x_1, \ldots, x_n).$$

(iv) Every symmetric gauge function is monotone on $\mathbb{R}^n_+$.

(v) If $x, y \in \mathbb{R}^n$ and $|x| \le |y|$, then $\Phi(x) \le \Phi(y)$ for every symmetric gauge function Φ.

(vi) If $x, y \in \mathbb{R}^n_+$, then $x \prec_w y$ if and only if $\Phi(x) \leq \Phi(y)$ for every symmetric gauge function Φ.

Problem II.5.12. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function such that $f(0) = 0$.

(i) Show that f is subadditive: $f(a+b) \leq f(a) + f(b)$ for all $a, b \in \mathbb{R}_+$.

(ii) Let $\Phi : \mathbb{R}^{2n}_+ \to \mathbb{R}_+$ be defined as

$$\Phi(x,y) = \sum_{j=1}^{n} f(x_j) + \sum_{j=1}^{n} f(y_j), \quad x, y \in \mathbb{R}^n_+.$$

Then Φ is Schur-concave.

(iii) Note that for $x, y \in \mathbb{R}^n_+$

$$(x,y) \prec (x+y, 0) \text{ in } \mathbb{R}^{2n}_+.$$

(iv) From (ii) and (iii) conclude that the function

$$F(x) = \sum_{j=1}^{n} f(|x_j|)$$

is subadditive on $\mathbb{R}^n$.

(v) Special examples lead to the following inequalities for vectors $x, y \in \mathbb{R}^n$:

$$\sum_{j=1}^{n} |x_j + y_j|^p \leq \sum_{j=1}^{n} |x_j|^p + \sum_{j=1}^{n} |y_j|^p, \quad 0 < p \leq 1.$$

$$\sum_{j=1}^{n} \frac{|x_j + y_j|}{1 + |x_j + y_j|} \leq \sum_{j=1}^{n} \frac{|x_j|}{1 + |x_j|} + \sum_{j=1}^{n} \frac{|y_j|}{1 + |y_j|}.$$

$$\sum_{j=1}^{n} \log\,(1 + |x_j + y_j|) \leq \sum_{j=1}^{n} \log(1 + |x_j|) + \sum_{j=1}^{n} \log(1 + |y_j|).$$

Problem II.5.13. Show that a map $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is lattice superadditive if and only if

$$\varphi(x_1 + \delta_1, x_2 - \delta_2) + \varphi(x_1 - \delta_1, x_2 + \delta_2)$$
$$\leq \varphi(x_1 + \delta_1, x_2 + \delta_2) + \varphi(x_1 - \delta_1, x_2 - \delta_2)$$

for all (x_1, x_2) and for all $\delta_1, \delta_2 \geq 0$. If φ is twice differentiable, this is equivalent to

$$0 \leq \frac{\partial^2 \varphi(x_1, x_2)}{\partial x_1 \partial x_2}.$$

Problem II.5.14. Let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ be a monotone increasing lattice superadditive function, and let f be a monotone increasing and convex function from $\mathbb{R}$ to $\mathbb{R}$. Show that if φ and f are twice differentiable, then the composition $f \circ \varphi$ is monotone and lattice superadditive. When $\varphi(s,t) = s+t$ show that this is also true if f is monotone decreasing. These statements are also true without any differentiability assumptions.

Problem II.5.15. For $x, y \in \mathbb{R}^n_+$

$$-\log(x^{\downarrow} + y^{\uparrow}) \prec_w -\log(x+y) \prec_w -\log(x^{\downarrow} + y^{\downarrow})$$

$$\log(x^{\downarrow} \cdot y^{\uparrow}) \prec_w \log(x \cdot y) \prec_w \log(x^{\downarrow} \cdot y^{\downarrow}).$$

From the first of these relations it follows that

$$\prod_{j=1}^{n}(x_j^{\downarrow} + y_j^{\downarrow}) \le \prod_{j=1}^{n}(x_j + y_j) \le \prod_{j=1}^{n}(x_j^{\downarrow} + y_j^{\uparrow}).$$

Problem II.5.16. Let x, y, u be vectors in $\mathbb{R}^n$ all having their coordinates in decreasing order. Show that

(i) $\langle x, u\rangle \le \langle y, u\rangle$ if $x \prec y$,

(ii) $\langle x, u\rangle \le \langle y, u\rangle$ if $x \prec_w y$ and $u \in \mathbb{R}^n_+$.

In particular, this means that if $x, y \in \mathbb{R}^n, x \prec_w y$, and $u \in \mathbb{R}^n_+$, then

$$(x_1^{\downarrow}u_1^{\downarrow}, \ldots, x_n^{\downarrow}u_n^{\downarrow}) \prec_w (y_1^{\downarrow}u_1^{\downarrow}, \ldots, y_n^{\downarrow}u_n^{\downarrow}).$$

[Use Theorem II.3.14 or the telescopic summation identity

$$\sum_{j=1}^{k} a_j b_j = \sum_{j=1}^{k}(a_j - a_{j+1})(b_1 + \cdots + b_j),$$

where $a_j, b_j,\ 1 \le j \le k$, are any numbers and $a_{k+1} = 0$.]

II.6 Notes and References

Many of the results of this chapter can be found in the classic *Inequalities* by G.H. Hardy, J.E. Littlewood, and G. Polya, Cambridge University Press, 1934, which gave the first systematic treatment of this theme. The more recent treatise *Inequalities: Theory of Majorization and Its Applications* by A.W. Marshall and I. Olkin, Academic Press, 1979, is a much more detailed and exhaustive text devoted entirely to the study of majorisation.

It is an invaluable resource on this topic. For the reader who wants a quicker introduction to the essentials of majorisation and its applications in linear algebra, the survey article *Majorization, doubly stochastic matrices and comparison of eigenvalues* by T. Ando, Linear Algebra and Its Applications, 118(1989) 163-248, is undoubtedly the ideal course. Our presentation is strongly influenced by this article from which we have freely borrowed.

The distance $d(\lambda, \mu)$ introduced in Problem II.5.9 is commonly employed in the study of variation of roots of polynomials and eigenvalues of matrices since these are known with no preferred ordering. See Chapter 6. The result of Problem II.5.10 is due to L. Elsner, C. Johnson, J. Ross, and J. Schönheim, *On a generalised matching problem arising in estimating the eigenvalue variation of two matrices*, European J. Combinatorics, 4(1983) 133-136.

Several of the theorems in this chapter have converses. For illustration we mention two of these.

Schur's Theorem (II.14) has a converse; it says that if d and λ are real vectors with $d \prec \lambda$, then there exists a Hermitian matrix A whose diagonal entries are the components of d and whose eigenvalues are the components of λ.

Weyl's Majorant Theorem (II.3.6) has a converse; it says that if $\lambda_1, \ldots, \lambda_n$ are complex numbers and $s_1, \ldots, s_n$ are positive real numbers ordered as $|\lambda_1| \geq \cdots \geq |\lambda_n|$ and $s_1 \geq \cdots \geq s_n$, and if

$$\prod_{j=1}^{k} |\lambda_j| \leq \prod_{j=1}^{k} s_j \quad \text{for} \quad 1 \leq k \leq n,$$

$$\prod_{j=1}^{n} |\lambda_j| = \prod_{j=1}^{n} s_j,$$

then there exists an $n \times n$ matrix A whose eigenvalues are $\lambda_1, \ldots, \lambda_n$ and singular values $s_1, \ldots, s_n$.

For more such theorems, see the book by Marshall and Olkin cited above.

Two results very close to those in II.3.16-II.3.21 and II.5.6-II.5.8 are given below.

M. Marcus and L. Lopes, *Inequalities for symmetric functions and Hermitian matrices*, Canad. J. Math., 9(1957) 305-312, showed that the map $\Phi : \mathbb{R}^n_+ \to \mathbb{R}$ given by $\Phi(x) = (S_k(x))^{1/k}$ is Schur-concave for $1 \leq k \leq n$. Using this they showed that for positive matrices A, B

$$[\text{tr} \wedge^k (A + B)]^{1/k} \geq [\text{tr} \wedge^k A]^{1/k} + [\text{tr} \wedge^k B]^{1/k}. \qquad \text{(II.44)}$$

This can also be expressed by saying that the map $A \to (\text{tr} \ \wedge^k A)^{1/k}$ is concave on the set of positive matrices. For $k = n$, this reduces to the statement

$$[\det(A + B)]^{1/n} \geq [\det A]^{1/n} + [\det B]^{1/n},$$

which is the Minkowski determinant inequality.

E.H. Lieb, *Convex trace functions and the Wigner-Yanase-Dyson conjecture*, Advances in Math., 11(1973) 267-288, proved some striking operator inequalities in connection with the W.-Y.-D. conjecture on the concavity of entropy in quantum mechanics. These were proved by different techniques and extended in other directions by T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26(1979) 203-241. One consequence of these results is the inequality

$$\wedge^k (A+B)^{1/k} \geq \wedge^k A^{1/k} + \wedge^k B^{1/k}, \tag{II.45}$$

for all positive matrices A, B and for all $k = 1, 2, \ldots, n$. In particular, this implies that

$$\text{tr } \wedge^k (A+B)^{1/k} \geq \text{tr } \wedge^k A^{1/k} + \text{tr } \wedge^k B^{1/k}.$$

When $k = n$, this reduces to the Minkowski determinant inequality. Some of these inequalities are proved in Chapter 9.

III

Variational Principles for Eigenvalues

In this chapter we will study inequalities that are used for localising the spectrum of a Hermitian operator. Such results are motivated by several interrelated considerations. It is not always easy to calculate the eigenvalues of an operator. However, in many scientific problems it is enough to know that the eigenvalues lie in some specified intervals. Such information is provided by the inequalities derived here. While the functional dependence of the eigenvalues on an operator is quite complicated, several interesting relationships between the eigenvalues of two operators A, B and those of their sum $A + B$ are known. These relations are consequences of variational principles. When the operator B is small in comparison to A, then $A + B$ is considered as a perturbation of A or an approximation to A. The inequalities of this chapter then lead to perturbation bounds or error bounds.

Many of the results of this chapter lead to generalisations, or analogues, or open problems in other settings discussed in later chapters.

III.1 The Minimax Principle for Eigenvalues

The following notation will be used throughout this chapter. If A, B are Hermitian operators, we will write their spectral resolutions as $Au_j = \alpha_j u_j, Bv_j = \beta_j v_j, 1 \leq j \leq n$, always assuming that the eigenvectors u_j and the eigenvectors v_j are orthonormal and that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$. When the dependence of the eigenvalues on the operator is to be emphasized, we will write $\lambda^{\downarrow}(A)$ for the vector with com-

ponents $\lambda_1^\downarrow(A), \ldots, \lambda_n^\downarrow(A)$, where $\lambda_j^\downarrow(A)$ are arranged in decreasing order; i.e., $\lambda_j^\downarrow(A) = \alpha_j$. Similarly, $\lambda^\uparrow(A)$ will denote the vector with components $\lambda_j^\uparrow(A)$ where $\lambda_j^\uparrow(A) = \alpha_{n-j+1}, 1 \le j \le n$.

Theorem III.1.1 *(Poincaré's Inequality) Let A be a Hermitian operator on $\mathcal{H}$ and let $\mathcal{M}$ be any k-dimensional subspace of $\mathcal{H}$. Then there exist unit vectors x, y in $\mathcal{M}$ such that $\langle x, Ax\rangle \le \lambda_k^\downarrow(A)$ and $\langle y, Ay\rangle \ge \lambda_k^\uparrow(A)$.*

Proof. Let $\mathcal{N}$ be the subspace spanned by the eigenvectors $u_k, \ldots, u_n$ of A corresponding to the eigenvalues $\lambda_k^\downarrow(A), \ldots, \lambda_n^\downarrow(A)$. Then

$$\dim \mathcal{M} + \dim \mathcal{N} = n + 1,$$

and hence the intersection of $\mathcal{M}$ and $\mathcal{N}$ is nontrivial. Pick up a unit vector x in $\mathcal{M} \cap \mathcal{N}$. Then we can write $x = \sum_{j=k}^{n} \xi_j u_j$, where $\sum_{j=k}^{n} |\xi_j|^2 = 1$. Hence,

$$\langle x, Ax\rangle = \sum_{j=k}^{n} |\xi_j|^2 \lambda_j^\downarrow(A) \le \sum_{j=k}^{n} |\xi_j|^2 \lambda_k^\downarrow(A) = \lambda_k^\downarrow(A).$$

This proves the first statement. The second can be obtained by applying this to the operator $-A$ instead of A. Equally well, one can repeat the argument, applying it to the given k-dimensional space $\mathcal{M}$ and the $(n - k + 1)$-dimensional space spanned by $u_1, u_2, \ldots, u_{n-k+1}$. ■

Corollary III.1.2 *(The Minimax Principle) Let A be a Hermitian operator on $\mathcal{H}$. Then*

$$\begin{aligned} \lambda_k^\downarrow(A) &= \max_{\substack{\mathcal{M} \subset \mathcal{H} \\ \dim \mathcal{M} = k}} \min_{\substack{x \in \mathcal{M} \\ \|x\| = 1}} \langle x, Ax\rangle \\ &= \min_{\substack{\mathcal{M} \subset \mathcal{H} \\ \dim \mathcal{M} = n-k+1}} \max_{\substack{x \in \mathcal{M} \\ \|x\| = 1}} \langle x, Ax\rangle. \end{aligned}$$

Proof. By Poincaré's inequality, if $\mathcal{M}$ is any k-dimensional subspace of $\mathcal{H}$, then $\min_x \langle x, Ax\rangle \le \lambda_k^\downarrow(A)$, where x varies over unit vectors in $\mathcal{M}$. But if $\mathcal{M}$ is the span of $\{u_1, \ldots, u_k\}$, then this last inequality becomes an equality. That proves the first statement. The second can be obtained from the first by applying it to $-A$ instead of A. ■

This minimax principle is sometimes called the **Courant-Fischer-Weyl minimax principle**.

Exercise III.1.3 *In the proof of the minimax principle we made a particular choice of $\mathcal{M}$. This choice is not always unique. For example, if $\lambda_k^\downarrow(A) = \lambda_{k+1}^\downarrow(A)$, there would be a whole 1-parameter family of such subspaces obtained by choosing different eigenvectors of A belonging to $\lambda_k^\downarrow(A)$.*

This is not surprising. More surprising, perhaps even shocking, is the fact that we could have $\lambda_k^{\downarrow}(A) = \min\{\langle x, Ax\rangle : x \in \mathcal{M}, ||x|| = 1\}$, *even for a* k*-dimensional subspace that is not spanned by eigenvectors of* A. *Find an example where this happens. (There is a simple example.)*

Exercise III.1.4 *In the proof of Theorem III.1.1 we used a basic principle of linear algebra:*

$$\begin{aligned}\dim(\mathcal{M}_1 \cap \mathcal{M}_2) &= \dim \mathcal{M}_1 + \dim \mathcal{M}_2 - \dim(\mathcal{M}_1 + \mathcal{M}_2) \\ &\geq \dim \mathcal{M}_1 + \dim \mathcal{M}_2 - n,\end{aligned}$$

for any two subspaces $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *of an* n*-dimensional vector space. Derive the corresponding inequality for an intersection of three subspaces.*

An equivalent formulation of the Poincaré inequality is in terms of compressions. Recall that if V is an isometry of a Hilbert space $\mathcal{M}$ into $\mathcal{H}$, then the **compression** of A by V is defined to be the operator $B = V^*AV$. Usually we suppose that $\mathcal{M}$ is a subspace of $\mathcal{H}$ and V is the injection map. Then A has a block-matrix representation in which B is the northwest corner entry:

$$A = \begin{pmatrix} B & \star \\ \star & \star \end{pmatrix}.$$

We say that B is the compression of A **to** the subspace $\mathcal{M}$.

Corollary III.1.5 *(Cauchy's Interlacing Theorem) Let* A *be a Hermitian operator on* $\mathcal{H}$, *and let* B *be its compression to an* $(n-k)$*-dimensional subspace* $\mathcal{N}$. *Then for* $j = 1, 2, \ldots, n-k$

$$\lambda_j^{\downarrow}(A) \geq \lambda_j^{\downarrow}(B) \geq \lambda_{j+k}^{\downarrow}(A). \tag{III.1}$$

Proof. For any j, let $\mathcal{M}$ be the span of the eigenvectors $v_1, \ldots, v_j$ of B corresponding to its eigenvalues $\lambda_1^{\downarrow}(B), \ldots, \lambda_j^{\downarrow}(B)$. Then $\langle x, Bx\rangle = \langle x, Ax\rangle$ for all $x \in \mathcal{M}$. Hence,

$$\lambda_j^{\downarrow}(B) = \min_{\substack{x \in \mathcal{M} \\ ||x||=1}} \langle x, Bx\rangle = \min_{\substack{x \in \mathcal{M} \\ ||x||=1}} \langle x, Ax\rangle \leq \lambda_j^{\downarrow}(A).$$

This proves the first assertion in (III.1).

Now apply this to $-A$ and its compression $-B$ to the given subspace $\mathcal{N}$. Note that

$$-\lambda_i^{\downarrow}(A) = \lambda_i^{\uparrow}(-A) = \lambda_{n-i+1}^{\downarrow}(-A) \quad \text{for all } 1 \leq i \leq n,$$

and

$$-\lambda_j^{\downarrow}(B) = \lambda_j^{\uparrow}(-B) = \lambda_{(n-k)-j+1}^{\downarrow}(-B) \quad \text{for all } 1 \leq j \leq n-k.$$

Choose $i = j+k$. Then the first inequality yields $-\lambda_j^\downarrow(B) \leq -\lambda_{j+k}^\downarrow(B)$, which is the second inequality in (III.1). ■

The above inequalities look especially nice when B is the compression of A to an $(n-1)$-dimensional subspace: then they say that

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \cdots \geq \beta_{n-1} \geq \alpha_n. \tag{III.2}$$

This explains why this is called an interlacing theorem.

Exercise III.1.6 *The Poincaré inequality, the minimax principle, and the interlacing theorem can be derived from each other. Find an independent proof for each of them using Exercise III.1.4. (This "dimension-counting" for intersections of subspaces will be used in later sections too.)*

Exercise III.1.7 *Let B be the compression of a Hermitian operator A to an $(n-1)$-dimensional space $\mathcal{M}$. If, for some k, the space $\mathcal{M}$ contains the vectors $u_1, \ldots, u_k$, then $\beta_j = \alpha_j$ for $1 \leq j \leq k$. If $\mathcal{M}$ contains $u_k, \ldots, u_n$, then $\alpha_j = \beta_{j-1}$ for $k \leq j \leq n$.*

Exercise III.1.8 *(i) Let A_n be the $n \times n$ tridiagonal matrix with entries $a_{ii} = 2\cos\theta$ for all i, $a_{ij} = 1$ if $|i-j| = 1$, and $a_{ij} = 0$ otherwise. The determinant of A_n is* $\sin(n+1)\theta/\sin\theta$.

(ii) Show that the eigenvalues of A_n are given by $2(\cos\theta + \cos\frac{j\pi}{n+1})$, $1 \leq j \leq n$.

(iii) The special case when $a_{ii} = -2$ for all i arises in Rayleigh's finite-dimensional approximation to the differential equation of a vibrating string. In this case the eigenvalues of A_n are

$$\lambda_j^\downarrow(A_n) = -4\sin^2\frac{j\pi}{2(n+1)}, \quad 1 \leq j \leq n.$$

(iv) Note that, for each $k < n$, the matrix A_{n-k} is a compression of A_n. This example provides a striking illustration of Cauchy's interlacing theorem.

It is illuminating to think of the variational characterisation of eigenvalues as a solution of a variational problem in analysis. If A is a Hermitian operator on $\mathbb{R}^n$, the search for the top eigenvalue of A is just the problem of maximising the function $F(x) = x^*Ax$ subject to the constraint that the function $G(x) = x^*x$ has the fixed value 1. The extremum must occur at a critical point, and using Lagrange multipliers the condition for a point x to be critical is $\nabla F(x) = \lambda \nabla G(x)$, which becomes $Ax = \lambda x$. Our earlier arguments got to the extremum problem from the algebraic eigenvalue problem, and this argument has gone the other way.

If additional constraints are imposed, the maximum can only decrease. Confining x to an $(n-k)$-dimensional subspace is equivalent to imposing

k linearly independent linear constraints on it. These can be expressed as $H_j(x) = 0$, where $H_j(x) = w_j^* x$ and the vectors $w_j, 1 \le j \le k$ are linearly independent. Introducing additional Lagrange multipliers μ_j, the condition for a critical point is now $\nabla F(x) = \lambda \nabla G(x) + \sum_j \mu_j \nabla H_j(x)$; i.e., $Ax - \lambda x$ is no longer required to be 0 but merely to be a linear combination of the w_j. Look at this in block-matrix terms. Our space has been decomposed into a direct sum of a space $\mathcal{N}$ and its orthogonal complement which is spanned by $\{w_1, \ldots, w_k\}$. Relative to this direct sum decomposition we can write

$$A = \begin{pmatrix} B & C \\ C^* & D \end{pmatrix}.$$

Our vector x is now constrained to be in $\mathcal{N}$, and the requirement for it to be a critical point is that $(A - \lambda I)\binom{x}{0}$ lies in $\mathcal{N}^\perp$. This is exactly requiring x to be an eigenvector of the compression B.

If two interlacing sets of real numbers are given, they can be realised as the eigenvalues of a Hermitian matrix and one of its compressions. This is a converse to one of the theorems proved above:

Theorem III.1.9 *Let $\alpha_j, 1 \le j \le n$, and $\beta_i, 1 \le i \le n-1$, be real numbers such that*

$$\alpha_1 \ge \beta_1 \ge \alpha_2 \ge \cdots \ge \beta_{n-1} \ge \alpha_n.$$

Then there exists a compression of the diagonal matrix $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ having $\beta_i, 1 \le i \le n-1$, as its eigenvalues.

Proof. Let $Au_j = \alpha_j u_j$; then $\{u_j\}$ constitute the standard orthonormal basis in $\mathbb{C}^n$. There is a one-to-one correspondence between $(n-1)$-dimensional orthogonal projection operators and unit vectors given by $P = I - zz^*$. Each unit vector, in turn, is completely characterised by its coordinates ζ_j with respect to the basis u_j. We have $z = \sum \zeta_j u_j = \sum (u_j^* z) u_j, \sum |\zeta_j|^2 = 1$. We will find conditions on the numbers ζ_j so that, for the corresponding orthoprojector $P = I - zz^*$, the compression of A to the range of P has eigenvalues β_i.

Since PAP is a Hermitian operator of rank $n-1$, we must have

$$\prod_{i=1}^{n-1} (\lambda - \beta_i) = \mathrm{tr} \ \wedge^{n-1} [P(\lambda I - A)P].$$

If E_j are the projectors defined as $E_j = I - u_j u_j^*$, then

$$\wedge^{n-1}(\lambda I - A) = \sum_{j=1}^{n} \prod_{k \ne j} (\lambda - \alpha_k) \wedge^{n-1} E_j.$$

Using the result of Problem I.6.9 one sees that

$$\wedge^{n-1} P \cdot \wedge^{n-1} E_j \cdot \wedge^{n-1} P = |\zeta_j|^2 \wedge^{n-1} P.$$

Since rank $\wedge^{n-1}P = 1$, the above three relations give

$$\prod_{i=1}^{n-1}(\lambda-\beta_i) = \sum_{j=1}^{n}|\zeta_j|^2[\prod_{k\neq j}(\lambda-\alpha_k)], \tag{III.3}$$

an identity between polynomials of degree $n-1$, which the ζ_j must satisfy if B has spectrum $\{\beta_i\}$.

We will show that the interlacing inequalities between α_j and β_i ensure that we can find ζ_j satisfying (III.3) and $\sum_{j=1}^{n}|\zeta_j|^2 = 1$. We may assume, without loss of generality, that the α_j are distinct. Put

$$\gamma_j = \frac{\prod_{i=1}^{n-1}(\alpha_j-\beta_i)}{\prod_{k\neq j}(\alpha_j-\alpha_k)}, \quad 1\leq j\leq n. \tag{III.4}$$

The interlacing property ensures that all γ_j are nonnegative. Now choose ζ_j to be any complex numbers with $|\zeta_j|^2=\gamma_j$. Then the equation (III.3) is satisfied for the values $\lambda=\alpha_j, 1\leq j\leq n$, and hence it is satisfied for all λ. Comparing the leading coefficients of the two sides of (III.3), we see that $\sum_j|\zeta_j|^2 = 1$. This completes the proof. ■

III.2 Weyl's Inequalities

Several relations between eigenvalues of Hermitian matrices A, B, and $A+B$ can be obtained using the ideas of the previous section. Most of these results were first proved by H. Weyl.

Theorem III.2.1 *Let A, B be $n\times n$ Hermitian matrices. Then,*

$$\lambda_j^{\downarrow}(A+B) \leq \lambda_i^{\downarrow}(A)+\lambda_{j-i+1}^{\downarrow}(B) \quad \text{for } i\leq j, \tag{III.5}$$

$$\lambda_j^{\downarrow}(A+B) \geq \lambda_i^{\downarrow}(A)+\lambda_{j-i+n}^{\downarrow}(B) \quad \text{for } i\geq j. \tag{III.6}$$

Proof. Let u_j, v_j, and w_j denote the eigenvectors of A, B, and $A+B$ respectively, corresponding to their eigenvalues in decreasing order. Let $i\leq j$. Consider the three subspaces spanned by $\{w_1,\ldots,w_j\}$, $\{u_i,\ldots,u_n\}$, and $\{v_{j-i+1},\ldots,v_n\}$ respectively. These have dimensions $j, n-i+1$, and $n-j+i$, and hence by Exercise III.1.4 they have a nontrivial intersection. Let x be a unit vector in their intersection. Then

$$\lambda_j^{\downarrow}(A+B) \leq \langle x,(A+B)x\rangle = \langle x,Ax\rangle+\langle x,Bx\rangle \leq \lambda_i^{\downarrow}(A)+\lambda_{j-i+1}^{\downarrow}(B).$$

This proves (III.5). If A and B in this inequality are replaced by $-A$ and $-B$, we get (III.6). ■

Corollary III.2.2 *For each* $j = 1, 2, \ldots, n$,

$$\lambda_j^{\downarrow}(A) + \lambda_n^{\downarrow}(B) \leq \lambda_j^{\downarrow}(A+B) \leq \lambda_j^{\downarrow}(A) + \lambda_1^{\downarrow}(B). \tag{III.7}$$

Proof. Put $i = j$ in the above inequalities. ■

It is customary to state these and related results as perturbation theorems, whereby B is a perturbation of A; that is $B = A + H$. In many of the applications H is small and the object is to give bounds for the distance of $\lambda(B)$ from $\lambda(A)$ in terms of $H = B - A$.

Corollary III.2.3 *(Weyl's Monotonicity Theorem) If H is positive, then*

$$\lambda_j^{\downarrow}(A+H) \geq \lambda_j^{\downarrow}(A) \quad \textit{for all } j.$$

Proof. By the preceding corollary, $\lambda_j^{\downarrow}(A+H) \geq \lambda_j^{\downarrow}(A) + \lambda_n^{\downarrow}(H)$, but all the eigenvalues of H are nonnegative. Alternately, note that $\langle x, (A+H)x\rangle \geq \langle x, Ax\rangle$ for all x and use the minimax principal. ■

Exercise III.2.4 *If H is positive and has rank k, then*

$$\lambda_j^{\downarrow}(A+H) \geq \lambda_j^{\downarrow}(A) \geq \lambda_{j+k}^{\downarrow}(A+H) \quad \textit{for } j = 1, 2, \ldots, n-k.$$

This is analogous to Cauchy's interlacing theorem.

Exercise III.2.5 *Let H be any Hermitian matrix. Then*

$$\lambda_j^{\downarrow}(A) - ||H|| \leq \lambda_j^{\downarrow}(A+H) \leq \lambda_j^{\downarrow}(A) + ||H||.$$

This can be restated as:

Corollary III.2.6 *(Weyl's Perturbation Theorem) Let A and B be Hermitian matrices. Then*

$$\max_j |\lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B)| \leq ||A - B||.$$

Exercise III.2.7 *For Hermitian matrices A, B, we have*

$$||A - B|| \leq \max_j |\lambda_j^{\downarrow}(A) - \lambda_j^{\uparrow}(B)|.$$

It is useful to have another formulation of the above two inequalities, which will be in conformity with more general results proved later.

We will denote by Eig A a diagonal matrix whose diagonal entries are the eigenvalues of A. If these are arranged in decreasing order, we write this matrix as $\text{Eig}^{\downarrow}(A)$; if in increasing order as $\text{Eig}^{\uparrow}(A)$. The results of Corollary III.2.6 and Exercise III.2.7 can then be stated as

Theorem III.2.8 *For any two Hermitian matrices A, B,*

$$||\text{Eig}^{\downarrow}(A) - \text{Eig}^{\downarrow}(B)|| \leq ||A - B|| \leq ||\text{Eig}^{\downarrow}(A) - \text{Eig}^{\uparrow}(B)||.$$

Weyl's inequality (III.5) is equivalent to an inequality due to Aronszajn connecting the eigenvalues of a Hermitian matrix to those of any two complementary principal submatrices. For this let us rewrite (III.5) as

$$\lambda^{\downarrow}_{i+j-1}(A+B) \leq \lambda^{\downarrow}_i(A) + \lambda^{\downarrow}_j(B), \tag{III.8}$$

for all indices i, j such that $i + j - 1 \leq n$.

Theorem III.2.9 *(Aronszajn's Inequality) Let C be an $n \times n$ Hermitian matrix partitioned as*

$$C = \begin{pmatrix} A & X \\ X^* & B \end{pmatrix},$$

where A is a $k \times k$ matrix. Let the eigenvalues of A, B, and C be $\alpha_1 \geq \cdots \geq \alpha_k, \beta_1 \geq \cdots \geq \beta_{n-k}$, and $\gamma_1 \geq \cdots \geq \gamma_n$, respectively. Then

$$\gamma_{i+j-1} + \gamma_n \leq \alpha_i + \beta_j \quad \textit{for all } i, j \textit{ with } i + j - 1 \leq n. \tag{III.9}$$

Proof. First assume that $\gamma_n = 0$. Then C is a positive matrix. Hence $C = D^*D$ for some matrix D. Partition D as $D = (D_1 \; D_2)$, where D_1 has k columns. Then

$$C = \begin{pmatrix} A & X \\ X^* & B \end{pmatrix} = \begin{pmatrix} D_1^*D_1 & D_1^*D_2 \\ D_2^*D_1 & D_2^*D_2 \end{pmatrix}.$$

Note that $DD^* = D_1D_1^* + D_2D_2^*$. Now the nonzero eigenvalues of the matrix $C = D^*D$ are the same as those of DD^*. The same is true for the matrices $A = D_1^*D_1$ and $D_1D_1^*$, and also for the matrices $B = D_2^*D_2$ and $D_2D_2^*$. Hence, using Weyl's inequality (III.8) we get (III.9) in this special case.

If $\gamma_n \neq 0$, subtract $\gamma_n I$ from C. Then all eigenvalues of A, B, and C are translated by $-\gamma_n$. By the special case considered above we have

$$\gamma_{i+j-1} - \gamma_n \leq (\alpha_i - \gamma_n) + (\beta_j - \gamma_n),$$

which is the same as (III.9). ■

We have derived Aronszajn's inequality from Weyl's inequality. But the argument above can be reversed. Let A, B be $n \times n$ Hermitian matrices and let $C = A + B$. Let the eigenvalues of these matrices be $\alpha_1 \geq \cdots \geq \alpha_n, \beta_1 \geq \cdots \geq \beta_n$, and $\gamma_1 \geq \cdots \geq \gamma_n$, respectively. We want to prove that $\gamma_{i+j-1} \leq \alpha_i + \beta_j$. This is the same as $\gamma_{i+j-1} - (\alpha_n + \beta_n) \leq (\alpha_i - \alpha_n) + (\beta_j - \beta_n)$. Hence, we can assume, without loss of generality, that both A and B are positive. Then $A = D_1^*D_1$ and $B = D_2^*D_2$ for some matrices D_1, D_2. Hence,

$$C = D_1^*D_1 + D_2^*D_2 = (D_1^* \quad D_2^*) \begin{pmatrix} D_1 \\ D_2 \end{pmatrix}.$$

Consider the $2n \times 2n$ matrix

$$E = \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} (D_1^* \;\; D_2^*) = \begin{pmatrix} D_1 D_1^* & D_1 D_2^* \\ D_2 D_1^* & D_2 D_2^* \end{pmatrix}.$$

Then the eigenvalues of E are the eigenvalues of C together with n zeroes. Aronszajn's inequality for the partitioned matrix E then gives Weyl's inequality (III.8).

By this procedure, several linear inequalities for the eigenvalues of a sum of Hermitian matrices can be transformed to those for the eigenvalues of block Hermitian matrices, and vice versa.

III.3 Wielandt's Minimax Principle

The minimax principle (Corollary III.1.2) gives an extremal characterisation for each eigenvalue α_j of a Hermitian matrix A. Ky Fan's maximum principle (Problem I.6.15 and Exercise II.1.13) provides an extremal characterisation for the sum $\alpha_1 + \cdots + \alpha_k$ of the top k eigenvalues of A. In this section we will prove a deeper result due to Wielandt that subsumes both these principles by providing an extremal representation of any sum $\alpha_{i_1} + \cdots + \alpha_{i_k}$. The proof involves a more elaborate dimension-counting for intersections of subspaces than was needed earlier.

We will denote by $V+W$ the vector sum of two vector spaces V and W, by $V - W$ any linear complement of a space W in V, and by span $\{v_1, \ldots, v_k\}$ the linear span of vectors $v_1, \ldots, v_k$.

Lemma III.3.1 *Let $W_1 \supset W_2 \supset \cdots \supset W_k$ be a decreasing chain of vector spaces with* $\dim W_j \geq k-j+1$. *Let $w_j, 1 \leq j \leq k-1$, be linearly independent vectors such that $w_j \in W_j$, and let U be their linear span. Then there exists a nonzero vector u in $W_1 - U$ such that the space $U +$* span $\{u\}$ *has a basis $v_1, \ldots, v_k$ with $v_j \in W_j, 1 \leq j \leq k$.*

Proof. This will be proved by induction on k. The statement is easily verified when $k = 2$. Assume that it is true for a chain consisting of $k-1$ spaces. Let $w_1, \ldots, w_{k-1}$ be the given vectors and U their linear span. Let S be the linear span of $w_2, \ldots, w_{k-1}$. Apply the induction hypothesis to the chain $W_2 \supset \cdots \supset W_k$ to pick up a vector v in $W_2 - S$ such that the space $S + \text{span}\{v\}$ is equal to $\text{span}\{v_2, \ldots, v_k\}$ for some linearly independent vectors $v_j \in W_j, j = 2, \ldots, k$. This vector v may or may not be in the space U. We will consider the two possibilities. Suppose $v \in U$. Then $U = S + \text{span}\{v\}$ because U is $(k-1)$-dimensional and S is $(k-2)$-dimensional. Since $\dim W_1 \geq k$, there exists a nonzero vector u in $W_1 - U$. Then $u, v_2, \ldots, v_k$ form a basis for $U + \text{span}\{u\}$. Put $u = v_1$. All requirements are now met. Suppose $v \notin U$. Then $w_1 \notin S + \text{span}\{v\}$, for if w_1 were a linear combination of $w_2, \ldots, w_{k-1}$ and v, then v would be a linear combination of

$w_1, w_2, \ldots, w_{k-1}$ and hence be an element of U. So, $\text{span}\{w_1, v_2, \ldots, v_k\}$ is a k-dimensional space that must, therefore, be $U + \text{span}\{v\}$. Now $w_1 \in W_1$ and $v_j \in W_j, j = 2, \ldots, k$. Again all requirements are met. ∎

Theorem III.3.2 *Let $V_1 \subset V_2 \subset \cdots \subset V_k$ be linear subspaces of an n-dimensional vector space V, with $\dim V_j = i_j, 1 \leq i_1 < i_2 < \cdots < i_k \leq n$. Let $W_1 \supset W_2 \supset \cdots \supset W_k$ be subspaces of V, with $\dim W_j = n - i_j + 1 = \operatorname{codim} V_j + 1$. Then there exist linearly independent vectors $v_j \in V_j, 1 \leq j \leq k$, and linearly independent vectors $w_j \in W_j, 1 \leq j \leq k$, such that*

$$\text{span}\{v_1, \ldots, v_k\} = \text{span}\{w_1, \ldots, w_k\}.$$

Proof. When $k = 1$ the statement is obviously true. (We have used this repeatedly in the earlier sections.) The general case will be proved by induction on k. So, let us assume that the theorem has been proved for $k - 1$ pairs of subspaces. By the induction hypothesis choose $v_j \in V_j$ and $w_j \in W_j, 1 \leq j \leq k-1$, two sets of linearly independent vectors having the same linear span U. Note that U is a subspace of V_k.

For $j = 1, \ldots, k$, let $S_j = W_j \cap V_k$. Then note that

$$\begin{aligned} n &\geq \dim W_j + \dim V_k - \dim S_j \\ &= (n - i_j + 1) + i_k - \dim S_j. \end{aligned}$$

Hence,

$$\dim S_j \geq i_k - i_j + 1 \geq k - j + 1.$$

Note that $S_1 \supset S_2 \supset \cdots \supset S_k$ are subspaces of V_k and $w_j \in S_j$ for $j = 1, 2, \ldots, k-1$. Hence, by Lemma III.3.1 there exists a vector u in $S_1 - U$ such that the space $U + \text{span}\{u\}$ has a basis $u_1, \ldots, u_k$, where $u_j \in S_j \subset W_j, j = 1, 2, \ldots, k$. But $U + \text{span}\{u\}$ is also the linear span of $v_1, \ldots, v_{k-1}$ and u. Put $v_k = u$. Then $v_j \in V_j, j = 1, 2, \ldots, k$, and they span the same space as the u_j. ∎

Exercise III.3.3 *If V is a Hilbert space, the vectors v_j and w_j in the statement of the above theorem can be chosen to be orthonormal.*

Proposition III.3.4 *Let A be a Hermitian operator on $\mathcal{H}$ with eigenvectors u_j belonging to eigenvalues $\lambda_j^\downarrow(A), j = 1, 2, \ldots, n$.*

(i) Let $\mathcal{V}_j = \text{span}\{u_1, \ldots, u_j\}, 1 \leq j \leq n$. Given indices $1 \leq i_1 < \cdots < i_k \leq n$, choose orthonormal vectors x_{i_j} from the spaces $\mathcal{V}_{i_j}, j = 1, \ldots, k$. Let $\mathcal{V}$ be the span of these vectors, and let $A_{\mathcal{V}}$ be the compression of A to the space $\mathcal{V}$. Then

$$\lambda_j^\downarrow(A_{\mathcal{V}}) \geq \lambda_{i_j}^\downarrow(A) \quad \textit{for} \quad j = 1, \ldots, k.$$

(ii) Let $\mathcal{W}_j = \text{span}\{u_j, \ldots, u_n\}, 1 \leq j \leq n$. Choose orthonormal vectors x_{i_j} from the spaces $\mathcal{W}_{i_j}, j = 1, \ldots, k$. Let $\mathcal{W}$ be the span of these vectors

and $A_{\mathcal{W}}$ the compression of A to $\mathcal{W}$. Then

$$\lambda_j^{\downarrow}(A_{\mathcal{W}}) \leq \lambda_{i_j}^{\downarrow}(A) \quad \textit{for} \quad j = 1, \ldots, k.$$

Proof. Let $y_1, \ldots, y_k$ be the eigenvectors of $A_{\mathcal{V}}$ belonging to its eigenvalues $\lambda_1^{\downarrow}(A_{\mathcal{V}}), \ldots, \lambda_k^{\downarrow}(A_{\mathcal{V}})$. Fix $j, 1 \leq j \leq k$, and in the space $\mathcal{V}$ consider the spaces spanned by $\{x_{i_1}, \ldots, x_{i_j}\}$ and $\{y_j, \ldots, y_k\}$, respectively. The dimensions of these two spaces add up to $k+1$, while the space $\mathcal{V}$ is k-dimensional. Hence there exists a unit vector u in the intersection of these two spaces. For this vector we have

$$\lambda_j^{\downarrow}(A_{\mathcal{V}}) \geq \langle u, A_{\mathcal{V}} u\rangle = \langle u, Au\rangle \geq \lambda_{i_j}^{\downarrow}(A).$$

This proves (i). The statement (ii) has exactly the same proof. ■

Theorem III.3.5 *(Wielandt's Minimax Principle) Let A be a Hermitian operator on an n-dimensional space $\mathcal{H}$. Then for any indices $1 \leq i_1 < \cdots < i_k \leq n$ we have*

$$\begin{aligned}\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A) &= \max_{\substack{\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_k \\ \dim \mathcal{M}_j = i_j}} \ \min_{\substack{x_j \in \mathcal{M}_j \\ x_j \text{ orthonormal}}} \sum_{j=1}^{k} \langle x_j, Ax_j\rangle \\ &= \min_{\substack{\mathcal{N}_1 \supset \cdots \supset \mathcal{N}_k \\ \dim \mathcal{N}_j = n - i_j + 1}} \ \max_{\substack{x_j \in \mathcal{N}_j \\ x_j \text{ orthonormal}}} \sum_{j=1}^{k} \langle x_j, Ax_j\rangle.\end{aligned}$$

Proof. We will prove the first statement; the second has a similar proof. Let $\mathcal{V}_{i_j} = \text{span}\{u_1, \ldots, u_{i_j}\}$, where, as before, the u_j are eigenvectors of A corresponding to $\lambda_j^{\downarrow}(A)$. For any unit vector x in $\mathcal{V}_{i_j}$, $\langle x, Ax\rangle \geq \lambda_{i_j}^{\downarrow}(A)$. So, if $x_j \in \mathcal{V}_{i_j}$ are orthonormal vectors, then

$$\sum_{j=1}^{k} \langle x_j, Ax_j\rangle \geq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A).$$

Since x_j were quite arbitrary, we have

$$\min_{\substack{x_j \in \mathcal{V}_{i_j} \\ x_j \text{ orthonormal}}} \sum_{j=1}^{k} \langle x_j, Ax_j\rangle \geq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A).$$

Hence, the desired result will be achieved if we prove that given any subspaces $\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_k$ with $\dim \mathcal{M}_j = i_j$ we can find orthonormal vectors $x_j \in \mathcal{M}_j$ such that

$$\sum_{j=1}^{k} \langle x_j, Ax_j\rangle \leq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A).$$

Let $\mathcal{N}_j = \mathcal{W}_{i_j} = \text{span}\{v_{i_j}, \ldots, v_n\}, j = 1, 2, \ldots, k$. These spaces were considered in Proposition III.3.4(ii). We have $\mathcal{N}_1 \supset \mathcal{N}_2 \cdots \supset \mathcal{N}_k$ and $\dim \mathcal{N}_j = n - i_j + 1$. Hence, by Theorem III.3.2 and Exercise III.3.3 there exist orthonormal vectors $x_j \in \mathcal{M}_j$ and orthonormal vectors $y_j \in \mathcal{N}_j$ such that

$$\text{span}\{x_1, \ldots, x_k\} = \text{span}\{y_1, \ldots, y_k\} = \mathcal{W}, \quad \text{say.}$$

By Proposition III.3.4 (ii), $\lambda_j^{\downarrow}(A_{\mathcal{W}}) \leq \lambda_{i_j}^{\downarrow}(A)$ for $j = 1, 2, \ldots, k$. Hence,

$$\begin{aligned} \sum_{j=1}^{k} \langle x_j, Ax_j \rangle &= \sum_{j=1}^{k} \langle x_j, A_{\mathcal{W}} x_j \rangle = \text{tr}\, A_{\mathcal{W}} \\ &= \sum_{j=1}^{k} \lambda_j^{\downarrow}(A_{\mathcal{W}}) \leq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A). \end{aligned}$$

This is what we wanted to prove. ■

Exercise III.3.6 *Note that*

$$\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A) = \sum_{j=1}^{k} \langle u_{i_j}, Au_{i_j} \rangle.$$

We have seen that the maximum in the first assertion of Theorem III.3.5 is attained when $\mathcal{M}_j = \mathcal{V}_{i_j} = \text{span}\{u_1, \ldots, u_{i_j}\}, j = 1, \ldots, k$, *and with this choice the minimum is attained for* $x_j = u_{i_j}, j = 1, \ldots, k$. *Are there other choices of subspaces and vectors for which these extrema are attained? (See Exercise III.1.3.)*

Exercise III.3.7 *Let* $[a, b]$ *be an interval containing all eigenvalues of* A *and let* $\Phi(t_1, \ldots, t_k)$ *be any real valued function on* $[a, b] \times \cdots \times [a, b]$ *that is monotone in each variable and permutation-invariant. Show that for each choice of indices* $1 \leq i_1 < \cdots < i_k \leq n$,

$$\begin{aligned} &\Phi\left(\lambda_{i_1}^{\downarrow}(A), \ldots, \lambda_{i_k}^{\downarrow}(A)\right) \\ &= \max_{\substack{\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_k \\ \dim \mathcal{M}_j = i_j}} \min_{\substack{\mathcal{W} = \text{span}\{x_1, \ldots, x_k\} \\ x_j \in \mathcal{M}_j, x_j \text{ orthonormal}}} \Phi\left(\lambda_1^{\downarrow}(A_{\mathcal{W}}), \ldots, \lambda_k^{\downarrow}(A_{\mathcal{W}})\right), \end{aligned}$$

where $A_{\mathcal{W}}$ *is the compression of* A *to the space* $\mathcal{W}$. *In Theorem III.3.5 we have proved the special case of this with* $\Phi(t_1, \ldots, t_k) = t_1 + \cdots + t_k$.

III.4 Lidskii's Theorems

One important application of Wielandt's minimax principle is in proving a theorem of Lidskii giving a relationship between eigenvalues of Hermitian

matrices A, B and $A+B$. This is quite like our derivation of some of the results in Section III.2 from those in Section III.1.

Theorem III.4.1 *Let A, B be Hermitian matrices. Then for any choice of indices $1 \leq i_1 < \cdots < i_k \leq n$,*

$$\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A+B) \leq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A) + \sum_{j=1}^{k} \lambda_{j}^{\downarrow}(B). \tag{III.10}$$

Proof. By Theorem III.3.5 there exist subspaces $\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_k$, with $\dim \mathcal{M}_j = i_j$ such that

$$\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A+B) = \min_{\substack{x_j \in \mathcal{M}_j \\ x_j \text{ orthonormal}}} \sum_{j=1}^{k} \langle x_j, (A+B)x_j \rangle.$$

By Ky Fan's maximum principle

$$\sum_{j=1}^{k} \langle x_j, Bx_j \rangle \leq \sum_{j=1}^{k} \lambda_{j}^{\downarrow}(B),$$

for any choice of orthonormal vectors $x_1, \ldots, x_k$. The above two relations imply that

$$\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A+B) \leq \min_{\substack{x_j \in \mathcal{M}_j \\ x_j \text{ orthonormal}}} \sum_{j=1}^{k} \langle x_j, Ax_j \rangle + \sum_{j=1}^{k} \langle x_j, Bx_j \rangle.$$

Now, using Theorem III.3.5 once again, it can be concluded that the first term on the right-hand side of the above inequality is dominated by $\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A)$. ■

Corollary III.4.2 *If A, B are Hermitian matrices, then the eigenvalues of A, B, and $A+B$ satisfy the following majorisation relation*

$$\lambda^{\downarrow}(A+B) - \lambda^{\downarrow}(A) \prec \lambda(B). \tag{III.11}$$

Exercise III.4.3 *(Lidskii's Theorem) The vector $\lambda^{\downarrow}(A+B)$ is in the convex hull of the vectors $\lambda^{\downarrow}(A) + P\lambda^{\downarrow}(B)$, where P varies over all permutation matrices. [This statement and those of Theorem III.4.1 and Corollary III.4.2 are, in fact, equivalent to each other.]*

Lidskii's Theorem can be proved without calling upon the more intricate Wielandt's principle. We will see several other proofs in this book, each highlighting a different viewpoint. The second proof given below is in the spirit of other results of this chapter.

Lidskii's Theorem (second proof). We will prove Theorem III.4.1 by induction on the dimension n. Its statement is trivial when $n = 1$. Assume it is true up to dimension $n-1$. When $k = n$, the inequality (III.10) needs no proof. So we may assume that $k < n$.

Let u_j, v_j, and w_j be the eigenvectors of A, B, and $A+B$ corresponding to their eigenvalues $\lambda_j^{\downarrow}(A)$, $\lambda_j^{\downarrow}(B)$, and $\lambda_j^{\downarrow}(A+B)$. We will consider three cases separately.

Case 1. $i_k < n$. Let $\mathcal{M} = \text{span}\{w_1, \ldots, w_{n-1}\}$ and let $A_{\mathcal{M}}$ be the compression of A to the space $\mathcal{M}$. Then, by the induction hypothesis

$$\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A_{\mathcal{M}} + B_{\mathcal{M}}) \leq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A_{\mathcal{M}}) + \sum_{j=1}^{k} \lambda_j^{\downarrow}(B_{\mathcal{M}}).$$

The inequality (III.10) follows from this by using the interlacing principle (III.2) and Exercise III.1.7.

Case 2. $1 < i_1$. Let $\mathcal{M} = \text{span}\{u_2, \ldots, u_n\}$. By the induction hypothesis

$$\sum_{j=1}^{k} \lambda_{i_j-1}^{\downarrow}(A_{\mathcal{M}} + B_{\mathcal{M}}) \leq \sum_{j=1}^{k} \lambda_{i_j-1}^{\downarrow}(A_{\mathcal{M}}) + \sum_{j=1}^{k} \lambda_j^{\downarrow}(B_{\mathcal{M}}).$$

Once again, the inequality (III.10) follows from this by using the interlacing principle and Exercise III.1.7.

Case 3. $i_1 = 1$. Given the indices $1 = i_1 < i_2 < \cdots < i_k \leq n$, pick up the indices $1 \leq \ell_1 < \ell_2 < \ldots < \ell_{n-k} < n$ such that the set $\{i_j : 1 \leq j \leq k\}$ is the complement of the set $\{n - \ell_j + 1 : 1 \leq j \leq n-k\}$ in the set $\{1, 2, \ldots, n\}$. These new indices now come under Case 1. Use (III.10) for this set of indices, but for matrices $-A$ and $-B$ in place of A, B. Then note that $\lambda_j^{\downarrow}(-A) = -\lambda_{n-j+1}^{\downarrow}(A)$ for all $1 \leq j \leq n$. This gives

$$\sum_{j=1}^{n-k} -\lambda_{n-\ell_j+1}^{\downarrow}(A+B) \leq \sum_{j=1}^{n-k} -\lambda_{n-\ell_j+1}^{\downarrow}(A) + \sum_{j=1}^{n-k} -\lambda_{n-j+1}^{\downarrow}(B).$$

Now add $\text{tr}(A+B)$ to both sides of the above inequality to get

$$\sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A+B) \leq \sum_{j=1}^{k} \lambda_{i_j}^{\downarrow}(A) + \sum_{j=1}^{k} \lambda_j^{\downarrow}(B).$$

This proves the theorem. ■

As in Section III.2, it is useful to interpret the above results as perturbation theorems. The following statement for Hermitian matrices A, B can be derived from (III.11) by changing variables:

$$\lambda^{\downarrow}(A) - \lambda^{\downarrow}(B) \prec \lambda(A-B) \prec \lambda^{\downarrow}(A) - \lambda^{\uparrow}(B). \qquad \text{(III.12)}$$

This can also be written as

$$\lambda^{\downarrow}(A) + \lambda^{\uparrow}(B) \prec \lambda(A+B) \prec \lambda^{\downarrow}(A) + \lambda^{\downarrow}(B). \tag{III.13}$$

In fact, the two right-hand majorisations are consequences of the weaker maximum principle of Ky Fan.

As a consequence of (III.12) we have:

Theorem III.4.4 *Let A, B be Hermitian matrices and let Φ be any symmetric gauge function on $\mathbb{R}^n$. Then*

$$\Phi\left(\lambda^{\downarrow}(A) - \lambda^{\downarrow}(B)\right) \leq \Phi\left(\lambda(A-B)\right) \leq \Phi\left(\lambda^{\downarrow}(A) - \lambda^{\uparrow}(B)\right).$$

Note that Weyl's perturbation theorem (Corollary III.2.6) and the inequality in Exercise III.2.7 are very special cases of this theorem.

The majorisations in (III.13) are significant generalisations of those in (II.35), which follow from these by restricting A, B to be diagonal matrices. Such "noncommutative" extensions exist for some other results; they are harder to prove. Some are given in this section; many more will occur later.

It is convenient to adopt the following notational shorthand. If x, y, z are n-vectors with nonnegative coordinates, we will write

$$\log x \prec_w \log y \quad \text{if} \quad \prod_{j=1}^{k} x_j^{\downarrow} \leq \prod_{j=1}^{k} y_j^{\downarrow}, \ \text{for } k = 1, \ldots, n; \tag{III.14}$$

$$\log x \prec \log y \quad \text{if} \quad \log x \prec_w \log y \quad \text{and} \quad \prod_{j=1}^{n} x_j^{\downarrow} = \prod_{j=1}^{n} y_j^{\downarrow}; \tag{III.15}$$

$$\log x - \log z \prec_w \log y \quad \text{if} \quad \prod_{j=1}^{k} x_{i_j} \leq \prod_{j=1}^{k} y_j \prod_{j=1}^{k} z_{i_j}, \tag{III.16}$$

for all indices $1 \leq i_1 < \cdots < i_k \leq n$. Note that we are allowing the possibility of zero coordinates in this notation.

Theorem III.4.5 *(Gel'fand-Naimark) Let A, B be any two operators on $\mathcal{H}$. Then the singular values of A, B and AB satisfy the majorisation*

$$\log s(AB) - \log s(B) \prec \log s(A). \tag{III.17}$$

Proof. We will use the result of Exercise III.3.7. Fix any index $k, 1 \leq k \leq n$. Choose any k orthonormal vectors $x_1, \ldots, x_k$, and let $\mathcal{W}$ be their linear span. Let $\Phi(t_1, \ldots, t_k) = t_1 t_2 \cdots t_k$. Express AB in its polar form $AB = UP$. Then, denoting by $T_{\mathcal{W}}$ the compression of an operator T to the subspace $\mathcal{W}$, we have

$$\begin{aligned}
\Phi\left(\lambda_1^2(P_{\mathcal{W}}), \ldots, \lambda_k^2(P_{\mathcal{W}})\right) &= |\det P_{\mathcal{W}}|^2 \\
&= |\det(\langle x_i, P_{\mathcal{W}} x_j \rangle)|^2 \\
&= |\det(\langle x_i, P x_j \rangle)|^2 \\
&= |\det(\langle A^* U x_i, B x_j \rangle)|^2.
\end{aligned}$$

Using Exercise I.5.7 we see that this is dominated by

$$\det(\langle A^*Ux_i, A^*Ux_j\rangle)\det(\langle Bx_i, Bx_j\rangle).$$

The second of these determinants is equal to $\det(B^*B)_{\mathcal{W}}$; the first is equal to $\det(AA^*)_{U\mathcal{W}}$ and by Corollary III.1.5 is dominated by $\prod_{j=1}^{k} s_j^2(A)$. Hence, we have

$$\Phi\left(\lambda_1^2(P_{\mathcal{W}}),\ldots,\lambda_k^2(P_{\mathcal{W}})\right) \le \det(B^*B)_{\mathcal{W}}\prod_{j=1}^{k} s_j^2(A)$$

$$= \Phi\left(\lambda_1(|B|^2_{\mathcal{W}}),\ldots,\lambda_k(|B|^2_{\mathcal{W}})\right)\prod_{j=1}^{k} s_j^2(A).$$

Now, using Exercise III.3.7, we can conclude that

$$\left(\prod_{j=1}^{k}\lambda_{i_j}^{\downarrow}(P)\right)^2 \le \prod_{j=1}^{k}\lambda_{i_j}^{\downarrow}(|B|^2)\prod_{j=1}^{k} s_j^2(A),$$

i.e.,

$$\prod_{j=1}^{k} s_{i_j}(AB) \le \prod_{j=1}^{k} s_{i_j}(B)\prod_{j=1}^{k} s_j(A), \tag{III.18}$$

for all $1 \le i_1 < \ldots < i_k \le n$. This, by definition, is what (III.17) says. ■

Remark. The statement

$$\prod_{j=1}^{k} s_j(AB) \le \prod_{j=1}^{k} s_j(A)\prod_{j=1}^{k} s_j(B), \tag{III.19}$$

which is a special case of (III.18), is easier to prove. It is just the statement $\| \wedge^k (AB)\| \le \| \wedge^k A\|\,\| \wedge^k B\|$. If we temporarily introduce the notation $s^{\downarrow}(A)$ and $s^{\uparrow}(A)$ for the vectors whose coordinates are the singular values of A arranged in decreasing order and in increasing order, respectively, then the inequalities (III.18) and (III.19) can be combined to yield

$$\log s^{\downarrow}(A) + \log s^{\uparrow}(B) \prec \log s(AB) \prec \log s^{\downarrow}(A) + \log s^{\downarrow}(B) \tag{III.20}$$

for any two matrices A, B. In conformity with our notation this is a symbolic representation of the inequalities

$$\prod_{j=1}^{k} s_{i_j}(A)\prod_{j=1}^{k} s_{n-i_j+1}(B) \le \prod_{j=1}^{k} s_j(AB) \le \prod_{j=1}^{k} s_j(A)\prod_{j=1}^{k} s_j(B)$$

for all $1 \le i_1 < \cdots < i_k \le n$. It is illuminating to compare this with the statement (III.13) for eigenvalues of Hermitian matrices.

Corollary III.4.6 *(Lidskii) Let A, B be two positive matrices. Then all eigenvalues of AB are nonnegative and*

$$\log \lambda^{\downarrow}(A) + \log \lambda^{\uparrow}(B) \prec \log \lambda(AB) \prec \log \lambda^{\downarrow}(A) + \log \lambda^{\downarrow}(B). \quad \text{(III.21)}$$

Proof. It is enough to prove this when B is invertible, since every positive matrix is a limit of such matrices. For invertible B we can write

$$AB = B^{-1/2}(B^{1/2}A\, B^{1/2})B^{1/2}.$$

Now $B^{1/2}A\, B^{1/2}$ is positive; hence the matrix AB, which is similar to it, has nonnegative eigenvalues. Now, from (III.20) we obtain

$$\begin{aligned}&\log \lambda^{\downarrow}(A^{1/2}) + \log \lambda^{\uparrow}(B^{1/2}) \\ &\quad \prec \log s(A^{1/2}B^{1/2}) \prec \log \lambda^{\downarrow}(A^{1/2}) + \log \lambda^{\downarrow}(B^{1/2}). \quad \text{(III.22)}\end{aligned}$$

But $s^2(A^{1/2}B^{1/2}) = \lambda^{\downarrow}(B^{1/2}AB^{1/2}) = \lambda^{\downarrow}(AB)$. So, the majorisations (III.21) follow from (III.22). ■

III.5 Eigenvalues of Real Parts and Singular Values

The Cartesian decomposition $A = \operatorname{Re} A + i \operatorname{Im} A$ of a matrix A associates with it two Hermitian matrices $\operatorname{Re} A = \frac{A+A^*}{2}$ and $\operatorname{Im} A = \frac{A-A^*}{2i}$. It is of interest to know relationships between the eigenvalues of these matrices, those of A, and the singular values of A.

Weyl's majorant theorem (Theorem II.3.6) provides one such relationship:

$$\log |\lambda(A)| \prec \log s(A).$$

Some others, whose proofs are in the same spirit as others in this chapter, are given below.

Proposition III.5.1 *(Fan-Hoffman) For every matrix A*

$$\lambda_j^{\downarrow}(\operatorname{Re} A) \le s_j(A) \quad \textit{for all} \quad j = 1, \ldots, n.$$

Proof. Let x_j be eigenvectors of $\operatorname{Re} A$ belonging to its eigenvalues $\lambda_j^{\downarrow}(\operatorname{Re} A)$ and y_j eigenvectors of $|A|$ belonging to its eigenvalues $s_j(A)$, $1 \le j \le n$. For each j consider the spaces $\operatorname{span}\{x_1, \ldots, x_j\}$ and $\operatorname{span}\{y_j, \ldots, y_n\}$. Their dimensions add up to $n+1$, so they have a nonzero intersection. If x is a unit vector in their intersection then

$$\begin{aligned}\lambda_j^{\downarrow}(\operatorname{Re} A) &\le \langle x, (\operatorname{Re} A)x\rangle = \operatorname{Re}\langle x, Ax\rangle \\ &\le |\langle x, Ax\rangle| \le \|Ax\| \\ &= \langle x, A^*Ax\rangle^{1/2} \le s_j(A).\end{aligned}$$

■

Exercise III.5.2 *(i) Let A be the 2×2 matrix $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$. Then $s_2(A) = 0$, but* Re A *has two nonzero eigenvalues. Hence the vector $|\lambda(\text{Re } A)|^{\downarrow}$ is not dominated by the vector $s(A)$.*

(ii) However, note that $|\lambda(\text{Re } A)| \prec_w s(A)$ for every matrix A. (Use the triangle inequality for Ky Fan norms.)

Proposition III.5.3 *(Ky Fan) For every matrix A we have*

$$\text{Re } \lambda(A) \prec \lambda(\text{Re } A).$$

Proof. Arrange the eigenvalues $\lambda_j(A)$ in such a way that

$$\text{Re } \lambda_1(A) \geq \text{Re } \lambda_2(A) \geq \cdots \geq \text{Re } \lambda_n(A).$$

Let $x_1, \ldots, x_n$ be an orthonormal Schur-basis for A such that $\lambda_j(A) = \langle x_j, Ax_j \rangle$. Then $\overline{\lambda_j(A)} = \langle x_j, A^* x_j \rangle$. Let $\mathcal{W} = \text{span}\{x_1, \ldots, x_k\}$. Then

$$\begin{aligned} \sum_{j=1}^{k} \text{Re } \lambda_j(A) &= \sum_{j=1}^{k} \langle x_j, (\text{Re } A) x_j \rangle = \text{tr } (\text{Re} A)_{\mathcal{W}} \\ &= \sum_{j=1}^{k} \lambda_j((\text{Re } A)_{\mathcal{W}}) \leq \sum_{j=1}^{k} \lambda_j^{\downarrow}(\text{Re } A). \end{aligned}$$

■

Exercise III.5.4 *Give another proof of Proposition III.5.3 using Schur's theorem (given in Exercise II.1.12).*

Exercise III.5.5 *Let X, Y be Hermitian matrices. Suppose that their eigenvalues can be indexed as $\lambda_j(X)$ and $\lambda_j(Y), 1 \leq j \leq n$, in such a way that $\lambda_j(X) \leq \lambda_j(Y)$ for all j. Then there exists a unitary U such that $X \leq U^* Y U$.*

(ii) For every matrix A there exists a unitary matrix U such that Re $A \leq U^* |A| U$.

An interesting consequence of Proposition III.5.1 is the following version of the triangle inequality for the matrix absolute value:

Theorem III.5.6 *(R.C. Thompson) Let A, B be any two matrices. Then there exist unitary matrices U, V such that*

$$|A + B| \leq U|A|U^* + V|B|V^*.$$

Proof. Let $A + B = W|A + B|$ be a polar decomposition of $A + B$. Then we can write

$$|A + B| = W^*(A + B) = \text{Re } W^*(A + B) = \text{Re } W^* A + \text{Re } W^* B.$$

Now use Exercise III.5.5(ii). ■

Exercise III.5.7 *(i) Find* 2×2 *matrices* A, B *such that the inequality* $|A+B| \le |A| + |B|$ *is false for them.*

(ii) Find 2×2 *matrices* A, B *for which there does not exist any unitary matrix* U *such that* $|A+B| \le U(|A|+|B|)U^*$.

III.6 Problems

Problem III.6.1. (The minimax principle for singular values) For any operator A on $\mathcal{H}$ we have

$$\begin{aligned} s_j(A) &= \max_{\mathcal{M}:\dim \mathcal{M}=j} \ \min_{x \in \mathcal{M}, ||x||=1} ||Ax|| \\ &= \min_{\mathcal{N}:\dim \mathcal{N}=n-j+1} \ \max_{x \in \mathcal{N}, ||x||=1} ||Ax|| \end{aligned}$$

for $1 \le j \le n$.

Problem III.6.2. Let A, B be any two operators. Then

$$s_j(AB) \le ||B|| s_j(A),$$

$$s_j(AB) \le ||A|| s_j(B)$$

for $1 \le j \le n$.

Problem III.6.3. For $j = 0, 1, \ldots, n$, let

$$\Re_j = \{T \in \mathcal{L}(H) : \text{rank}\, T \le j\}.$$

Show that for $j = 1, 2, \ldots, n$,

$$s_j(A) = \min_{T \in \Re_{j-1}} ||A - T||.$$

Problem III.6.4. Show that if A is any operator and H is any operator of rank k, then

$$s_j(A) \ge s_{j+k}(A+H), \quad j = 1, 2, \ldots, n-k.$$

Problem III.6.5. For any two operators A, B and any two indices i, j such that $i + j \le n + 1$, we have

$$s_{i+j-1}(A+B) \le s_i(A) + s_j(B)$$

$$s_{i+j-1}(AB) \le s_i(A) s_j(B).$$

Problem III.6.6. Show that for every operator A and for each $k = 1, 2, \ldots, n$, we have

$$\sum_{j=1}^{k} s_j(A) = \max \left| \sum_{j=1}^{k} \langle y_j, Ax_j \rangle \right|,$$

where the maximum is over all choices of orthonormal k-tuples $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$. This can also be written as

$$\sum_{j=1}^{k} s_j(A) = \max \left| \sum_{j=1}^{k} \langle x_j, UAx_j \rangle \right|,$$

where the maximum is taken over all choices of unitary operators U and orthonormal k-tuples $x_1, \ldots, x_k$. Note that for $k = 1$ this reduces to the statement

$$||A|| = \sum_{||x||=||y||=1} |\langle y, Ax \rangle|.$$

For $k = 1, 2, \ldots, n$, the above extremal representations can be used to give another proof of the fact that the expressions $||A||_{(k)} = \sum_{j=1}^{k} s_j(A)$ are norms. (See Exercise II.1.15.)

Problem III.6.7. Let $A = (a_{ij})$ be a Hermitian matrix. For each $i = 1, \ldots, n$, let

$$r_i = \left(\sum_{j \neq i} |a_{ij}|^2 \right)^{1/2}.$$

Show that each interval $[a_{ii} - r_i, a_{ii} + r_i]$ contains at least one eigenvalue of A.

Problem III.6.8. Let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ be the eigenvalues of a Hermitian matrix A. We have seen that the $n-1$ eigenvalues of any principal submatrix of A interlace with these numbers. If $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_{n-1}$ are the roots of the polynomial that is the derivative of the characteristic polynomial of A, then we have by Rolle's Theorem

$$\alpha_1 \geq \delta_1 \geq \alpha_2 \geq \cdots \geq \delta_{n-1} \geq \alpha_n.$$

Show that for each j there exists a principal submatrix B of A for which $\alpha_j \geq \lambda_j^{\downarrow}(B) \geq \delta_j$ and another principal submatrix C for which $\delta_j \geq \lambda_j^{\downarrow}(C) \geq \alpha_{j+1}$.

Problem III.6.9. Most of the results in this chapter gave descriptions of eigenvalues of a Hermitian operator in terms of the numbers $\langle x, Ax \rangle$

when x varies over unit vectors. Sometimes in computational problems an "approximate" eigenvalue λ and an "approximate" eigenvector x are already known. The number $\langle x, Ax\rangle$ can then be used to further refine this information.

For a given unit vector x, let $\rho = \langle x, Ax\rangle, \varepsilon = ||(A-\rho)x||$.

(i) Let (a, b) be an open interval that contains ρ but does not contain any eigenvalue of A. Show that

$$(b-\rho)(\rho-a) \leq \varepsilon^2.$$

(ii) Show that there exists an eigenvalue α of A such that $|\alpha - \rho| \leq \varepsilon$.

Problem III.6.10. Let ρ and ε be defined as in the above problem. Let (a, b) be an open interval that contains ρ and only one eigenvalue α of A. Then

$$\rho - \frac{\varepsilon^2}{\rho - a} \leq \alpha \leq \rho + \frac{\varepsilon^2}{b-\rho}.$$

This is called the **Kato-Temple inequality**. Note that if $\rho - a$ and $b - \rho$ are much larger than ε, then this improves the inequality in part (ii) of Problem III.6.9.

Problem III.6.11. Show that for every Hermitian matrix A

$$\sum_{j=1}^{k} \lambda_j^{\downarrow}(A) = \max_{UU^*=I_k} \operatorname{tr} UAU^*,$$

$$\sum_{j=1}^{k} \lambda_j^{\uparrow}(A) = \min_{UU^*=I_k} \operatorname{tr} UAU^*$$

for $1 \leq k \leq n$, where the extrema are taken over $k \times n$ matrices U that satisfy $UU^* = I_k$, I_k being the $k \times k$ identity matrix. Show that if A is positive, then

$$\prod_{j=1}^{k} \lambda_j^{\downarrow}(A) = \max_{UU^*=I_k} \det UAU^*,$$

$$\prod_{j=1}^{k} \lambda_j^{\uparrow}(A) = \min_{UU^*=I_k} \det UAU^*.$$

(See Problem I.6.15.)

Problem III.6.12. Let A, B be any matrices. Then

$$\sum_{j=1}^{n} s_j(A)s_j(B) = \sup_{U,V} |\operatorname{tr} UAVB| = \sup_{U,V} \operatorname{Re} \operatorname{tr} UAVB,$$

where U, V vary over all unitary matrices.

Problem III.6.13. (Perturbation theorem for singular values) Let A, B be any $n \times n$ matrices and let Φ be any symmetric gauge function on $\mathbb{R}^n$. Then

$$\Phi\left(s(A) - s(B)\right) \prec_w \Phi\left(s(A-B)\right).$$

In particular,

$$\max |s_j(A) - s_j(B)| \leq ||A - B||.$$

[Hint: See Theorem III.4.4 and Exercise II.1.15.]

Problem III.6.14. For positive matrices A, B show that

$$\lambda^{\downarrow}(A) \cdot \lambda^{\uparrow}(B) \prec \lambda(AB) \prec \lambda^{\downarrow}(A) \cdot \lambda^{\downarrow}(B).$$

For Hermitian matrices A, B show that

$$\langle \lambda^{\downarrow}(A), \lambda^{\uparrow}(B) \rangle \leq \operatorname{tr} AB \leq \langle \lambda^{\downarrow}(A), \lambda^{\downarrow}(B) \rangle.$$

(Compare these with (II.36) and (II.37).)

Problem III.6.15. Let A, B be Hermitian matrices. Use the second part of Problem III.6.14 to show that

$$||\mathrm{Eig}^{\downarrow} A - \mathrm{Eig}^{\downarrow} B||_2 \leq ||A - B||_2 \leq ||\mathrm{Eig}^{\downarrow} A - \mathrm{Eig}^{\uparrow} B||_2.$$

Note the analogy between this and Theorem III.2.8. (In Chapter IV we will see that both these results are true for a whole family of norms called unitarily invariant norms. This more general result is a consequence of Theorem III.4.4.)

III.7 Notes and References

As pointed out in Exercise III.1.6, many of the results in Sections III.1 and III.2 could be derived from each other. Hence, it seems fair to say that the variational principles for eigenvalues originated with A.L. Cauchy's interlacing theorem. A pertinent reference is *Sur l'équation á l'aide de laquelle on détermine les inégalités séculaires des mouvements des planétes*, 1829, in A.L. Cauchy, *Oeuvres Complétes (IIe Série)*, Volume 9, Gauthier-Villars.

The minimax principle was first stated by E. Fischer, *Über Quadratische Formen mit reellen Koeffizienten*, Monatsh. Math. Phys., 16 (1905) 234-249. The monotonicity principle and many of the results of Section III.2 were proved by H. Weyl in *Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen*, Math. Ann., 71 (1911)441-469. In a series of papers beginning with *Über die Eigenwerte bei den Differentialgleichungen der mathematischen Physik*, Math. Z., 7(1920) 1-57,

R. Courant exploited the full power of the minimax principle. Thus the principle is often described as the Courant-Fischer-Weyl principle.

As the titles of these papers suggest, the variational principles for eigenvalues were discovered in connections with problems of physics. One famous work where many of these were used is *The Theory of Sound* by Lord Rayleigh, reprinted by Dover in 1945. The modern applied mathematics classic *Methods of Mathematical Physics* by R. Courant and D. Hilbert, Wiley, 1953, is replete with applications of variational principles. For a still more recent source, see M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Volume 4, Academic Press, 1978. Of course, here most of the interest is in infinite-dimensional problems and consequently the results are much more complicated. The numerical analyst could turn to B.N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, 1980, and to G.W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, 1990.

The converse to the interlacing theorem given in Theorem III.1.9 was first proved in L. Mirsky, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc., 33 (1958) 14-21. We do not know whether the similar question for higher dimensional compressions has been answered. More precisely, let $\alpha_1 \geq \cdots \geq \alpha_n$, and $\beta_1 \geq \cdots \geq \beta_n$, be real numbers such that $\Sigma\alpha_j = \Sigma\beta_j$. What conditions must these numbers satisfy so that there exists an orthogonal projection P of rank k such that the matrix $A = \text{diag}\ (\alpha_1, \ldots, \alpha_n)$ when compressed to range P has eigenvalues $\beta_1, \ldots, \beta_k$ and when compressed to (range $P)^\perp$ has eigenvalues $\beta_{k+1}, \ldots, \beta_n$? (Theorem III.1.9 is the case $k = n-1$.)

Aronszajn's inequality appeared in N. Aronszajn, *Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues. I. Operators in a Hilbert space*, Proc. Nat. Acad. Sci. U.S.A., 34(1948) 474-480. The elegant proof of its equivalence to Weyl's inequality is due to H.W. Wielandt, *Topics in the Analytic Theory of Matrices*, mimeographed lecture notes, University of Wisconsin, 1967.

Theorem III.3.5 was proved in H.W. Wielandt, *An extremum property of sums of eigenvalues*, Proc. Amer. Math. Soc., 6 (1955) 106-110. The motivation for Wielandt was that he "did not succeed in completing the interesting sketch of a proof given by Lidskii" of the statement given in Exercise III.4.3. He noted that this is equivalent to what we have stated as Theorem III.4.1, and derived it from his new minimax principle. Interestingly, now several different proofs of Lidskii's Theorem are known. The second proof given in Section III.4 is due to M.F. Smiley, *Inequalities related to Lidskii's*, Proc. Amer. Math. Soc., 19 (1968) 1029-1034. We will see some other proofs later. However, Theorem III.3.5 is more general, has several other applications, and has led to a lot of research. An account of the earlier work on these questions may be found in A.R. Amir-Moez, *Extreme Properties of Linear Transformations and Geometry in Unitary Spaces*, Texas Tech. University, 1968, from which our treatment of Section III.3 has been adapted. An attempt to extend these ideas to infinite

dimensions was made in R.C. Riddell, *Minimax problems on Grassmann manifolds*, Advances in Math., 54 (1984) 107-199, where connections with differential geometry and some problems in quantum physics are also developed. The tower of subspaces occurring in Theorem III.3.5 suggests a connection with Schubert calculus in algebraic geometry. This connection is yet to be fully understood.

Lidskii's Theorem has an interesting history. It appeared first in V.B. Lidskii, *On the proper values of a sum and product of symmetric matrices*, Dokl. Akad. Nauk SSSR, 75 (1950) 769-772. It seems that Lidskii provided an elementary (matrix analytic) proof of the result which F. Berezin and I.M. Gel'fand had proved by more advanced (Lie theoretic) techniques in connection with their work that appeared later in *Some remarks on the theory of spherical functions on symmetric Riemannian manifolds*, Trudi Moscow Math. Ob., 5 (1956) 311-351. As mentioned above, difficulties with this "elementary" proof led Wielandt to the discovery of his minimax principle.

Among the several directions this work opened up, one led to the following question. What relations must three n-tuples of real numbers satisfy in order to be the eigenvalues of some Hermitian matrices A, B and $A+B$? Necessary conditions are given by Theorem III.4.1. Many more were discovered by others. A. Horn, *Eigenvalues of sums of Hermitian matrices*, Pacific J. Math., 12(1962) 225-242, derived necessary and sufficient conditions in the above problem for the case $n = 4$, and wrote down a set of conditions which he conjectured would be necessary and sufficient for $n > 4$. In a short paper *Spectral polyhedron of a sum of two Hermitian matrices*, Functional Analysis and Appl., 10 (1982) 76-77, B.V. Lidskii has sketched a "proof" establishing Horn's conjecture. This proof, however, needs a lot of details to be filled in; these have not yet been published by B.V. Lidskii (or anyone else).

When should a theorem be considered to be proved? For an interesting discussion of this question, see S. Smale, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (New Series), 4(1981) 1-36.

Theorem III.4.5 was proved in I.M. Gel'fand and M. Naimark, *The relation between the unitary representations of the complex unimodular group and its unitary subgroup*, Izv Akad. Nauk SSSR Ser. Mat. 14(1950) 239-260. Many of the questions concerning eigenvalues and singular values of sums and products were first framed in this paper. An excellent summary of these results can be found in A.S. Markus, *The eigen-and singular values of the sum and product of linear operators*, Russian Math. Surveys, 19 (1964) 92-120.

The structure of inequalities like (III.10) and (III.18) was carefully analysed in several papers by R.C. Thompson and his students. The asymmetric way in which A and B enter (III.10) is remedied by one of their inequalities,

which says

$$\sum_{j=1}^{k} \lambda^{\downarrow}_{i_j+p_j-j}(A+B) \leq \sum_{j=1}^{k} \lambda^{\downarrow}_{i_j}(A) + \sum_{j=1}^{k} \lambda^{\downarrow}_{p_j}(B)$$

for any indices $1 \leq i_1 < \cdots < i_k \leq n, 1 \leq p_1 < \cdots < p_k \leq n$, such that $i_k + p_k - k \leq n$. A similar generalisation of (III.18) has also been proved. References to this work may be found in the book by Marshall and Olkin cited in Chapter II.

Proposition III.5.1 is proved in K. Fan and A.J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955) 111-116.

Results of Proposition III.5.3, Problems III.6.5, III.6.6, III.6.11, and III.6.12 were first proved by Ky Fan in several papers. References to these may be found in I.C. Gohberg and M.G. Krein, *Introduction to the Theory of Linear Nonselfadjoint operators*, American Math. Society, 1969, and in the Marshall-Olkin book cited earlier.

The matrix triangle inequality (Theorem III.5.6) was proved in R.C. Thompson, *Convex and concave functions of singular values of matrix sums*, Pacific J. Math., 66 (1976) 285-290. An extension to infinite dimensions was attempted in C. Akemann, J. Anderson, and G. Pedersen, *Triangle inequalities in operator algebras*, Linear and Multilinear Algebra, 11(1982) 167-178. For operators A, B on an infinite-dimensional Hilbert space there exist isometries U, V such that

$$|A+B| \leq U|A|U^* + V|B|V^*.$$

Also, for each $\varepsilon > 0$ there exist unitaries U, V such that

$$|A+B| \leq U|A|U^* + V|B|V^* + \varepsilon I.$$

It is not known whether the ε part in the last statement is necessary.

Refinements of the interlacing principle such as the one in Problem III.6.8 have been obtained by several authors, including R.C. Thompson. See, for example, his paper *Principal submatrices II*, Linear Algebra Appl., 1(1968) 211-243.

One may wonder whether there are interlacing theorem, for singular values. There are, although they are a little different from the ones for eigenvalues. This is best understood if we extend the definition of singular values to rectangular matrices. Let A be an $m \times n$ matrix. Let $r = \min(m, n)$. The r numbers that are the common eigenvalues of $(A^*A)^{1/2}$ and $(AA^*)^{1/2}$ are called the singular values of A. (Sometimes a sequence of zeroes is added to make $\max(m, n)$ singular values in all.) Many of the results for singular values that we have proved can be carried over to this setting. See, e.g., the books by Horn and Johnson cited in Chapter I.

Let A be a rectangular matrix and let B be a matrix obtained by deleting any row or any column of A. Then the minimax principle can be used to prove that the singular values of A and B interlace. The reader should work this out, and see that when A is an $n \times n$ matrix and B a principal submatrix of order $n-1$ then this gives

$$\begin{array}{rcccl} s_1(A) & \geq & s_1(B) & \geq & s_3(A), \\ s_2(A) & \geq & s_2(B) & \geq & s_4(A), \\ \cdots\cdots & & \cdots\cdots & & \\ s_{n-2}(A) & \geq & s_{n-2}(B) & \geq & s_n(A), \\ s_{n-1}(A) & \geq & s_{n-1}(B) & \geq & 0. \end{array}$$

For more such results, see R.C. Thompson, *Principal submatrices IX*, Linear Algebra and Appl., 5(1972) 1-12.

Inequalities like the ones in Problems III.6.9 and III.6.10 are called "residual bounds" in the numerical analysis literature. For more such results, see the book by Parlett cited above, and F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, 1983. =Several refinements, extensions, and applications of these results in atomic physics are described in the book by Reed and Simon cited above.

The results of Theorem III.4.4 and Problem III.6.13 were noted by L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math., Oxford Ser. (2), 11(1960) 50-59. This paper contains a lucid survey of several related problems and has stimulated a lot of research. The inequalities in Problem III.6.15 were first stated in K. Löwner, *Über monotone Matrix functionen*, Math. Z., 38 (1934) 177-216.

Let $A = UP$ be a polar decomposition of A. Weyl's majorant theorem gives a relationship between the eigenvalues of A and those of P (the singular values of A). A relation between the eigenvalues of A and those of U was proved by A. Horn and R. Steinberg, *Eigenvalues of the unitary part of a matrix*, Pacific J. Math., 9(1959) 541-550. This is in the form of a majorisation between the arguments of the eigenvalues:

$$\arg \lambda(A) \prec \arg \lambda(U).$$

A theorem, very much like Theorems III.4.1 and III.4.5 was proved by A. Nudel'man and P. Svarcman, *The spectrum of a product of unitary matrices*, Uspehi Mat. Nauk, 13 (1958) 111-117. Let A, B be unitary matrices. Label the eigenvalues of A, B, and AB as $e^{i\alpha_1}, \ldots, e^{i\alpha_n}; e^{i\beta_1}, \ldots, e^{i\beta_n}$, and $e^{i\gamma_1}, \ldots, e^{i\gamma_n}$, respectively, in such a way that

$$2\pi > \alpha_1 \geq \cdots \geq \alpha_n \geq 0,$$

$$2\pi > \beta_1 \geq \cdots \geq \beta_n \geq 0,$$

$$2\pi > \gamma_1 \geq \cdots \geq \gamma_n \geq 0.$$

If $\alpha_1 + \beta_1 < 2\pi$, then for any choice of indices $1 \leq i_1 < \cdots < i_k \leq n$ we have

$$\sum_{j=1}^{k} \gamma_{i_j} \leq \sum_{j=1}^{k} \alpha_{i_j} + \sum_{j=1}^{k} \beta_j .$$

These inequalities can also be written in the form of a majorisation between n-vectors:

$$\gamma - \alpha \prec \beta .$$

For a generalisation in the same spirit as the one of inequalities (III.10) and (III.18) mentioned earlier, see R.C. Thompson, *On the eigenvalues of a product of unitary matrices*, Linear and Multilinear Algebra, 2(1974) 13-24.

IV

Symmetric Norms

In this chapter we study norms on the space of matrices that are invariant under multiplication by unitaries. Their properties are closely linked to those of symmetric gauge functions on $\mathbb{R}^n$. We also study norms that are invariant under unitary conjugations. Some of the inequalities proved in earlier chapters lead to inequalities involving these norms.

IV.1 Norms on $\mathbb{C}^n$

Let us begin by considering the familiar p-norms frequently used in analysis. For a vector $x = (x_1, \ldots, x_n)$ we define

$$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}, \quad 1 \leq p < \infty, \tag{IV.1}$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|. \tag{IV.2}$$

For each $1 \leq p \leq \infty$, $\|x\|_p$ defines a norm on $\mathbb{C}^n$. These are called the p-norms or the l_p-norms. The notation (IV.2) is justified because of the fact that

$$\|x\|_\infty = \lim_{p \to \infty} \|x\|_p. \tag{IV.3}$$

Some of the pleasant properties of this family of norms are

$$\|x\|_p = \| \, |x| \, \|_p \quad \text{for all } x \in \mathbb{C}^n, \tag{IV.4}$$

$$\|x\|_p \leq \|y\|_p \quad \text{if } |x| \leq |y|, \tag{IV.5}$$

$$\|x\|_p = \|Px\|_p \quad \text{for all } x \in \mathbb{C}^n, P \in S_n. \tag{IV.6}$$

(Recall the notations: $|x| = (|x_1|, \ldots, |x_n|)$, and $|x| \leq |y|$ if $|x_j| \leq |y_j|$ for $1 \leq j \leq n$. S_n is the set of permutation matrices.) A norm on $\mathbb{C}^n$ is called **gauge invariant** or **absolute** if it satisfies the condition (IV.4), **monotone** if it satisfies (IV.5), and **permutation invariant** or **symmetric** if it satisfies (IV.6). The first two of these conditions turn out to be equivalent:

Proposition IV.1.1 *A norm on $\mathbb{C}^n$ is gauge invariant if and only if it is monotone.*

Proof. Monotonicity clearly implies gauge invariance. Conversely, if a norm $\|\cdot\|$ is gauge invariant, then to show that it is monotone it is enough to show that $\|x\| \leq \|y\|$ whenever $x_j = t_j y_j$ for some real numbers $0 \leq t_j \leq 1, j = 1, 2, \ldots, n$. Further, it suffices to consider the special case when all t_j except one are equal to 1. But then

$$\begin{aligned}
&\|(y_1, \ldots, ty_k, \ldots, y_n)\| \\
&= \left\| \left(\frac{1+t}{2} y_1 + \frac{1-t}{2} y_1, \ldots, \frac{1+t}{2} y_k - \frac{1-t}{2} y_k, \ldots, \frac{1+t}{2} y_n + \frac{1-t}{2} y_n \right) \right\| \\
&\leq \frac{1+t}{2} \|(y_1, \ldots, y_n)\| + \frac{1-t}{2} \|(y_1, \ldots, -y_k, \ldots, y_n)\| \\
&= \|(y_1, \ldots, y_n)\|.
\end{aligned}$$

■

Example IV.1.2 *Consider the following norms on $\mathbb{R}^2$:*

(i) $\|x\| = |x_1| + |x_2| + |x_1 - x_2|$.

(ii) $\|x\| = |x_1| + |x_1 - x_2|$.

(iii) $\|x\| = 2|x_1| + |x_2|$.

The first of these is symmetric but not gauge invariant, the second is neither symmetric nor gauge invariant, while the third is not symmetric but is gauge invariant.

Norms that are both symmetric and gauge invariant are especially interesting. Before studying more examples and properties of such norms, let us make a few remarks.

Let $\mathbf{T}$ be the circle group; i.e., the multiplicative group of all complex numbers of modulus 1. Let $S_n o \mathbf{T}$ be the semidirect product of S_n and $\mathbf{T}$. In other words, this is the group of all $n \times n$ matrices that have exactly one nonzero entry on each row and each column, and this nonzero entry has modulus 1. We will call such matrices **complex permutation matrices**. Then a norm $\|\cdot\|$ on $\mathbb{C}^n$ is symmetric and gauge invariant if

$$\|x\| = \|Tx\| \quad \text{for all complex permutations } T. \tag{IV.7}$$

In other words, the group of (linear) isometries for $\|\cdot\|$ contains $S_n o \mathbf{T}$ as a subgroup. (Linear isometries for a norm $\|\cdot\|$ are those linear transformations on $\mathbb{C}^n$ that preserve $\|\cdot\|$.)

Exercise IV.1.3 *For the Euclidean norm $\|x\|_2 = (\sum |x_i|^2)^{1/2}$ the group of isometries is the group of all unitary matrices, which is much larger than the complex permutation group. Show that for each of the norms $\|x\|_1$ and $\|x\|_\infty$ the group of isometries is the complex permutation group.*

Note that gauge invariant norms on $\mathbb{C}^n$ are determined by those on $\mathbb{R}^n$. Symmetric gauge invariant norms on $\mathbb{R}^n$ are called symmetric gauge functions. We have come across them earlier (Example II.3.13). To repeat, a map $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ is called a **symmetric gauge function** if

(i) Φ is a norm,

(ii) $\Phi(Px) = \Phi(x)$ for all $x \in \mathbb{R}^n$ and $P \in S_n$,

(iii) $\Phi(\varepsilon_1 x_1, \ldots, \varepsilon_n x_n) = \Phi(x_1, \ldots, x_n)$ if $\varepsilon_j = \pm 1$.

In addition, we will always assume that Φ is normalised, so that

(iv) $\Phi(1, 0, \ldots, 0) = 1$.

The conditions (ii) and (iii) can be expressed together by saying that Φ is invariant under the group $S_n o \mathbf{Z_2}$ consisting of permutations and sign changes of the coordinates. Notice also that a symmetric gauge function is completely determined by its values on $\mathbb{R}^n_+$.

Example IV.1.4 *If the coordinates of x are arranged so that $|x_1| \geq |x_2| \geq \cdots \geq |x_n|$, then for each $k = 1, 2, \ldots, n$, the function*

$$\Phi_{(k)}(x) = \sum_{j=1}^{k} |x_j| \tag{IV.8}$$

is a symmetric gauge function. We will also use the notation $\|x\|_{(k)}$ for these. The parentheses are used to distinguish these norms from the p-norms defined earlier. Indeed, note that $\|x\|_{(1)} = \|x\|_\infty$ and $\|x\|_{(n)} = \|x\|_1$.

We have observed in Problem II.5.11 that these norms play a very distinguished role: if $\Phi_{(k)}(x) \leq \Phi_{(k)}(y)$ for all $k = 1, 2, \ldots, n$, then $\Phi(x) \leq \Phi(y)$ for every symmetric gauge function Φ. Thus an infinite family of norm inequalities follows from a finite one.

Proposition IV.1.5 *For each $k = 1, 2, \ldots, n$,*

$$\Phi_{(k)}(x) = \min\{\Phi_{(n)}(u) + k\Phi_{(1)}(v) \ : \ x = u + v\}. \tag{IV.9}$$

Proof. We may assume, without loss of generality, that $x \in \mathbb{R}^n_+$. If $x = u+v$, then $\Phi_{(k)}(x) \le \Phi_{(k)}(u) + \Phi_{(k)}(v) \le \Phi_{(n)}(u) + k\Phi_{(1)}(v)$. If we choose

$$\begin{aligned} u &= (x_1^\downarrow - x_k^\downarrow, x_2^\downarrow - x_k^\downarrow, \dots, x_k^\downarrow - x_k^\downarrow, 0, \dots, 0) \\ v &= (x_k^\downarrow, \dots, x_k^\downarrow, x_{k+1}^\downarrow, \dots, x_n^\downarrow), \end{aligned}$$

then

$$\begin{aligned} u+v &= x^\downarrow, \\ \Phi_{(n)}(u) &= \Phi_{(k)}(x) - kx_k^\downarrow, \\ \Phi_{(1)}(v) &= x_k^\downarrow, \end{aligned}$$

and the proposition follows. ■

We now derive some basic inequalities. If f is a convex function on an interval I and if $a_i, i = 1, 2, \dots, n$, are nonnegative real numbers such that $\sum_{i=1}^n a_i = 1$, then

$$f(\sum_{i=1}^n a_i t_i) \le \sum_{i=1}^n a_i f(t_i) \quad \text{for all } t_i \in I.$$

Applying this to the function $f(t) = -\log t$ on the interval $(0, \infty)$, one obtains the fundamental inequality

$$\prod_{i=1}^n t_i^{a_i} \le \sum_{i=1}^n a_i t_i \quad \text{if} \quad t_i \ge 0, a_i \ge 0, \sum a_i = 1. \tag{IV.10}$$

This is called the **(weighted) arithmetic-geometric mean inequality**. The special choice $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$ gives the usual arithmetic - geometric mean inequality

$$(\prod_{i=1}^n t_i)^{1/n} \le \frac{1}{n} \sum_{i=1}^n t_i \quad \text{if} \quad t_i \ge 0. \tag{IV.11}$$

Theorem IV.1.6 *Let p, q be real numbers with $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Let $x, y \in \mathbb{R}^n$. Then for every symmetric gauge function Φ*

$$\Phi(|x \cdot y|) \le [\Phi(|x|^p)]^{1/p} [\Phi(|y|^q)]^{1/q}. \tag{IV.12}$$

Proof. From the inequality (IV.10) one obtains

$$|x \cdot y| \le \frac{|x|^p}{p} + \frac{|y|^q}{q},$$

and hence

$$\Phi(|x \cdot y|) \leq \frac{1}{p}\Phi(|x|^p) + \frac{1}{q}\Phi(|y|^q). \tag{IV.13}$$

For $t > 0$, if we replace x, y by tx and $t^{-1}y$, then the left-hand side of (IV.13) does not change. Hence,

$$\Phi(|x \cdot y|) \leq \min_{t>0}\left[\frac{t^p}{p}\Phi(|x|^p) + \frac{1}{qt^q}\Phi(|y|^q)\right]. \tag{IV.14}$$

But, if

$$\varphi(t) = \frac{t^p}{p}a + \frac{1}{qt^q}b, \quad \text{where } t, a, b > 0,$$

then plain differentiation shows that

$$\min \varphi(t) = a^{1/p}b^{1/q}.$$

So, (IV.12) follows from (IV.14). ■

When $\Phi = \Phi_{(n)}$, (IV.12) reduces to the familiar **Hölder inequality**

$$\sum_{i=1}^{n} |x_i y_i| \leq (\sum_{i=1}^{n} |x_i|^p)^{1/p}(\sum_{i=1}^{n} |y_i|^q)^{1/q}.$$

We will refer to (IV.12) as the **Hölder inequality for symmetric gauge functions**. The special case $p = 2$ will be called the **Cauchy-Schwarz inequality for symmetric gauge functions.**

Exercise IV.1.7 *Let p, q, r be positive real numbers with $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$. Show that for every symmetric gauge function Φ we have*

$$[\Phi(|x \cdot y|^r)]^{1/r} \leq [\Phi(|x|^p)]^{1/p}[\Phi(|y|^q)]^{1/q}. \tag{IV.15}$$

Theorem IV.1.8 *Let Φ be any symmetric gauge function and let $p \geq 1$. Then for all $x, y \in \mathbb{R}^n$*

$$[\Phi(|x + y|^p)]^{1/p} \leq [\Phi(|x|^p)]^{1/p} + [\Phi(|y|^p)]^{1/p}. \tag{IV.16}$$

Proof. When $p = 1$, the inequality (IV.16) is a consequence of the triangle inequalities for the absolute value on $\mathbb{R}^n$ and for the norm Φ. Let $p > 1$. It is enough to consider the case $x \geq 0$, $y \geq 0$. Make this assumption and write

$$(x + y)^p = x \cdot (x + y)^{p-1} + y \cdot (x + y)^{p-1}.$$

Now, using the triangle inequality for Φ and Theorem IV.1.6, one obtains

$$\begin{aligned} \Phi((x+y)^p) &\leq \Phi(x \cdot (x+y)^{p-1}) + \Phi(y \cdot (x+y)^{p-1}) \\ &\leq [\Phi(x^p)]^{1/p}[\Phi((x+y)^{q(p-1)})]^{1/q} \\ &+ [\Phi(y^p)]^{1/p}[\Phi((x+y)^{q(p-1)})]^{1/q} \\ &= \{[\Phi(x^p)]^{1/p} + [\Phi(y^p)]^{1/p}\}[\Phi((x+y)^p)]^{1/q}, \end{aligned}$$

since $q(p-1) = p$. If we divide both sides of the above inequality by $[\Phi((x+y)^p)]^{1/q}$, we get (IV.16). ∎

Once again, when $\Phi = \Phi_{(n)}$ the inequality (IV.16) reduces to the familiar Minkowski inequality. So, we will call (IV.16) the **Minkowski inequality for symmetric gauge functions.**

Exercise IV.1.9 *Let Φ be a symmetric gauge function and let $p \geq 1$. Let*

$$\Phi^{(p)}(x) = [\Phi(|x|^p)]^{1/p}. \qquad \text{(IV.17)}$$

Show that $\Phi^{(p)}$ is also a symmetric gauge function.

Note that, if Φ_p is the family of ℓ_p-norms, then

$$\Phi_{p_1}^{(p_2)} = \Phi_{p_1 p_2} \quad \textit{for all } \; p_1, p_2 \geq 1, \qquad \text{(IV.18)}$$

and, if $\Phi_{(k)}$ is the norm defined by (IV.8), then

$$\Phi_{(k)}^{(p)}(x) = (\sum_{j=1}^{k} |x_j|^p)^{1/p}, \qquad \text{(IV.19)}$$

where the coordinates of x are arranged as $|x_1| \geq |x_2| \geq \cdots \geq |x_n|$.

Just as among the l_p-norms, the Euclidean norm has especially interesting properties, the norms $\Phi^{(2)}$ where Φ is any symmetric gauge function have some special interest. We will give these norms a name:

Definition IV.1.10 *Ψ is called a* **quadratic symmetric gauge function,** *or a* **Q-norm,** *if $\Psi = \Phi^{(2)}$ for some symmetric gauge function Φ. In other words,*

$$\Psi(x) = [\Phi(|x|^2)]^{1/2}. \qquad \text{(IV.20)}$$

Exercise IV.1.11 *(i) Show that an l_p-norm is a Q-norm if and only if $p \geq 2$.*

(ii) More generally, show that for each $k = 1, 2, \ldots, n$, $\Phi_{(k)}^{(p)}$ is a Q-norm if and only if $p \geq 2$.

Exercise IV.1.12 *We saw earlier that if $\Phi_{(k)}(x) \leq \Phi_{(k)}(y)$ for all $k = 1, 2, \ldots, n$, then $\Phi(x) \leq \Phi(y)$ for all symmetric gauge functions. Show that if $\Phi_{(k)}^{(2)}(x) \leq \Phi_{(k)}^{(2)}(y)$ for all $k = 1, 2, \ldots, n$, then $\Phi^{(2)}(x) \leq \Phi^{(2)}(y)$ for all symmetric gauge functions Φ; i.e., $\Psi(x) \leq \Psi(y)$ for all quadratic symmetric gauge functions.*

If Φ is a norm on $\mathbb{C}^n$, the **dual** of Φ is defined as

$$\Phi'(x) = \sup_{\Phi(y)=1} |\langle x, y\rangle|. \qquad \text{(IV.21)}$$

It is easy to see that Φ' is a norm. (In fact, Φ' is a norm even when Φ is a function on $\mathbb{C}^n$ that does not necessarily satisfy the triangle inequality that but meets the other requirements of a norm.)

Exercise IV.1.13 *If Φ is a symmetric gauge function then so is Φ'.*

Exercise IV.1.14 *Show that for any norm Φ*

$$|\langle x, y\rangle| \leq \Phi(x)\Phi'(y) \quad \text{for all } x, y. \tag{IV.22}$$

Exercise IV.1.15 *Let Φ_p be the l_p-norm, $1 \leq p \leq \infty$. Show that*

$$\Phi'_p = \Phi_q, \quad \text{where} \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{IV.23}$$

Exercise IV.1.16 *Let Φ and Ψ be two norms such that*

$$\Phi(x) \leq c\Psi(x) \text{ for all } x \text{ and for some } c > 0.$$

Show that

$$\Phi'(x) \geq c^{-1}\Psi'(x) \quad \text{for all } x.$$

We shall call a symmetric gauge function a **Q'-norm** if it is the dual of a Q-norm. The l_p-norms for $1 \leq p \leq 2$ are examples of Q'-norms.

Exercise IV.1.17 *(i) Let Φ be a norm such that $\Phi = \Phi'$. Then Φ must be the Euclidean norm.*

(ii) Let Φ be both a Q-norm and a Q'-norm. Then Φ must be the Euclidean norm. (Use Exercise IV.1.16 and the fact that every symmetric gauge function is bounded by the l_1-norm.)

Exercise IV.1.18 *For each $k = 1, 2, \ldots, n$, the dual of the norm $\Phi_{(k)}$ is given by*

$$\Phi'_{(k)}(x) = \max\left\{\Phi_{(1)}(x), \frac{1}{k}\Phi_{(n)}(x)\right\}. \tag{IV.24}$$

Prove this using Proposition IV.1.5 and Exercise IV.1.16.

Some ways of generating symmetric gauge functions are described in the following exercises.

Exercise IV.1.19 *Let $1 = \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0$. Given a symmetric gauge function Φ on $\mathbb{R}^n$, define*

$$\Psi(x) = \Phi(\alpha_1|x|_1^{\downarrow}, \ldots, \alpha_n|x|_n^{\downarrow}).$$

Then Ψ is a symmetric gauge function.

Exercise IV.1.20 *(i) Let Φ be a symmetric gauge function on $\mathbb{R}^n$. Let $m < n$. If $x \in \mathbb{R}^m$, let $\tilde{x} = (x_1, \ldots, x_m, 0, 0, \ldots, 0)$ and define $\Psi(x) = \Phi(\tilde{x})$. Then Ψ is a symmetric gauge function on $\mathbb{R}^m$.*

(ii) Conversely, given any symmetric gauge function Ψ on $\mathbb{R}^m$, if for $n > m$ we define $\Phi(x_1, \ldots, x_n) = \Psi(|x|_1^{\downarrow}, \ldots, |x|_m^{\downarrow})$, then Φ is a symmetric gauge function on $\mathbb{R}^n$.

IV.2 Unitarily Invariant Norms on Operators on $\mathbb{C}^n$

In this section, $\mathbb{C}^n$ will always stand for the Hilbert space $\mathbb{C}^n$ with inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\|\cdot\|$. (No subscript will be attached to this "standard" norm as was done in the previous Section.) If A is a linear operator on $\mathbb{C}^n$, we will denote by $\|A\|$ the **operator (bound) norm** of A defined as

$$\|A\| = \sup_{\|x\|=1} \|Ax\|. \tag{IV.25}$$

As before, we denote by $|A|$ the positive operator $(A^*A)^{1/2}$ and by $s(A)$ the vector whose coordinates are the singular values of A, arranged as $s_1(A) \geq s_2(A) \geq \cdots \geq s_n(A)$. We have

$$\|A\| = \|\, |A| \,\| = s_1(A). \tag{IV.26}$$

Now, if U, V are unitary operators on $\mathbb{C}^n$, then $|UAV| = V^*|A|V$ and hence

$$\|A\| = \|UAV\| \tag{IV.27}$$

for all unitary operators U, V. This last property is called **unitary invariance**. Several other norms have this property. These are frequently useful in analysis, and we will study them in some detail.

We will use the symbol $|||\cdot|||$ to mean a norm on $n \times n$ matrices that satisfies

$$|||UAV||| = |||A||| \tag{IV.28}$$

for all A and for unitary U, V. We will call such a norm a **unitarily invariant norm** on the space $\mathbf{M(n)}$ of $n \times n$ matrices. We will normalise such norms so that they all take the value 1 on the matrix $\mathrm{diag}(1, 0, \ldots, 0)$.

There is an intimate connection between these norms and symmetric gauge functions on $\mathbb{R}^n$; the link is provided by singular values.

Theorem IV.2.1 *Given a symmetric gauge function Φ on $\mathbb{R}^n$, define a function on $\mathbf{M(n)}$ as*

$$|||A|||_\Phi = \Phi(s(A)). \tag{IV.29}$$

Then this defines a unitarily invariant norm on $\mathbf{M(n)}$. Conversely, given any unitarily invariant norm $|||\cdot|||$ on $\mathbf{M(n)}$, define a function on $\mathbb{R}^n$ by

$$\Phi_{|||\cdot|||}(x) = |||\mathrm{diag}(x)|||, \tag{IV.30}$$

where $\mathrm{diag}\,(x)$ *is the diagonal matrix with entries $x_1, \ldots, x_n$ on its diagonal. Then this defines a symmetric gauge function on $\mathbb{R}^n$.*

Proof. Since $s(UAV) = s(A)$ for all unitary U, V, $|||\cdot|||_\Phi$ is unitarily invariant. We will prove that it obeys the triangle inequality — the other

conditions for it to be a norm are easy to verify. For this, recall the majorisation (II.18)

$$s(A+B) \prec_w s(A)+s(B) \quad \text{for all } A, B \in \mathbf{M(n)},$$

and then use the fact that Φ is strongly isotone and monotone. (See Example II.3.13 and Problem II.5.11.) To prove the converse, note that (IV.30) clearly gives a norm on $\mathbb{R}^n$. Since diagonal matrices of the form $\operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$ and permutation matrices are all unitary, this norm is absolute and permutation invariant, and hence it is a symmetric gauge function. ■

Symmetric gauge functions on $\mathbb{R}^n$ constructed in the preceding section thus lead to several examples of unitarily invariant norms on $\mathbf{M(n)}$. Two classes of such norms are specially important. The first is the class of **Schatten p-norms** defined as

$$\|A\|_p = \Phi_p(s(A)) = [\sum_{j=1}^{n} (s_j(A))^p]^{1/p}, \quad 1 \leq p < \infty, \tag{IV.31}$$

$$\|A\|_\infty = \Phi_\infty(s(A)) = s_1(A) = \|A\|. \tag{IV.32}$$

The second is the class of **Ky Fan k-norms** defined as

$$\|A\|_{(k)} = \sum_{j=1}^{k} s_j(A), \quad 1 \leq k \leq n. \tag{IV.33}$$

Among the p-norms, the ones for the values $p = 1, 2, \infty$, are used most often. As we have noted, $\|A\|_\infty$ is the same as the operator norm $\|A\|$ and the Ky Fan norm $\|A\|_{(1)}$. The norm $\|A\|_1$ is the same as $\|A\|_{(n)}$. This is equal to $\operatorname{tr}(|A|)$ and hence is called the **trace norm**, and is sometimes written as $\|A\|_{tr}$. The norm

$$\|A\|_2 = [\sum_{j=1}^{n} (s_j(A))^2]^{1/2} \tag{IV.34}$$

is also called the **Hilbert-Schmidt** norm or the **Frobenius norm** (and is sometimes written as $\|A\|_F$ for that reason). It will play a basic role in our analysis. For $A, B \in \mathbf{M(n)}$ let

$$\langle A, B\rangle = \operatorname{tr} A^* B. \tag{IV.35}$$

This defines an inner product on $\mathbf{M(n)}$ and the norm associated with this inner product is $\|A\|_2$, i.e. ,

$$\|A\|_2 = (\operatorname{tr} A^* A)^{1/2}. \tag{IV.36}$$

If the matrix A has entries a_{ij}, then

$$\|A\|_2 = (\sum_{i,j} |a_{ij}|^2)^{1/2}. \tag{IV.37}$$

Thus the norm $\|A\|_2$ is the Euclidean norm of the matrix A when it is thought of as an element of $\mathbb{C}^{n^2}$. This fact makes this norm easily computable and geometrically tractable.

The main importance of the Ky Fan norms lies in the following:

Theorem IV.2.2 *(Fan Dominance Theorem) Let A, B be two $n \times n$ matrices. If*

$$\|A\|_{(k)} \le \|B\|_{(k)} \quad \text{for } k = 1, 2, \ldots, n,$$

then

$$|||A||| \le |||B||| \quad \text{for all unitarily invariant norms.}$$

Proof. This is a consequence of the corresponding assertion about symmetric gauge functions. (See Example IV.1.4.) ■

Since $\Phi_{(1)}(x) \le \Phi(x) \le \Phi_{(n)}(x)$ for all $x \in \mathbb{R}^n$ and for all symmetric gauge functions Φ, we have

$$\|A\| \le |||A||| \le \|A\|_{(n)} = \|A\|_1 \tag{IV.38}$$

for all $A \in \mathbf{M(n)}$ and for all unitarily invariant norms $|||\cdot|||$.

Analogous to Proposition IV.1.5 we have

Proposition IV.2.3 *For each $k = 1, 2, \ldots, n$,*

$$\|A\|_{(k)} = \min\{\|B\|_{(n)} + k\|C\| : A = B + C\}. \tag{IV.39}$$

Proof. If $A = B + C$, then $\|A\|_{(k)} \le \|B\|_{(k)} + \|C\|_{(k)} \le \|B\|_{(n)} + k\|C\|$. Now let $s(A) = (s_1, \ldots, s_n)$ and choose unitary U, V so that

$$A = U[(\text{diag}(s_1, \ldots, s_n)]V.$$

Let

$$\begin{aligned} B &= U[\text{diag}(s_1 - s_k, s_2 - s_k, \ldots, s_k - s_k, 0, \ldots, 0)]V, \\ C &= U[\text{diag}(s_k, s_k, \ldots, s_k, s_{k+1}, \ldots, s_n)]V. \end{aligned}$$

Then

$$A = B + C,$$

$$\|B\|_{(n)} = \sum_{j=1}^{k} s_j - ks_k = \|A\|_{(k)} - ks_k,$$

$$\|C\| = s_k,$$

and

$$\|A\|_{(k)} = \|B\|_{(n)} + k\|C\|. \qquad \blacksquare$$

A norm ν on $\mathbf{M(n)}$ is called **symmetric** if for A, B, C in $\mathbf{M(n)}$

$$\nu(BAC) \le \|B\|\ \nu(A)\ \|C\|. \tag{IV.40}$$

Proposition IV.2.4 *A norm on* $\mathbf{M(n)}$ *is symmetric if and only if it is unitarily invariant.*

Proof. If ν is a symmetric norm, then for unitary U, V we have $\nu(UAV) \le \nu(A)$ and $\nu(A) = \nu(U^{-1}UAVV^{-1}) \le \nu(UAV)$. So, ν is unitarily invariant. Conversely, by Problem III.6.2, $s_j(BAC) \le \|B\|\ \|C\| s_j(A)$ for all $j = 1, 2, \ldots, n$. So, if Φ is any symmetric gauge function, then $\Phi(s(BAC)) \le \|B\|\ \|C\|\Phi(s(A))$ and hence the norm associated with Φ is symmetric. ∎

In particular, this implies that every unitarily invariant norm is **submultiplicative**:

$$|||AB||| \le |||A|||\ |||B||| \quad \text{for all } A, B.$$

Inequalities for sums and products of singular values of matrices, when combined with inequalities for symmetric gauge functions proved in Section IV.1, lead to interesting statements about unitarily invariant norms. This is illustrated below.

Theorem IV.2.5 *If* A, B *are* $n \times n$ *matrices, then*

$$s^r(AB) \prec_w s^r(A)s^r(B) \quad \text{for all}\, r > 0. \tag{IV.41}$$

Proof. If $\wedge^k A$ is the kth antisymmetric tensor product of A, then

$$\| \wedge^k A\| = s_1(\wedge^k A) = \prod_{j=1}^{k} s_j(A), \quad 1 \le k \le n.$$

Hence,

$$\begin{aligned} \prod_{j=1}^{k} s_j^r(AB) &= \| \wedge^k (AB)\|^r \le (\| \wedge^k A\|\ \| \wedge^k B\|)^r \\ &= \prod_{j=1}^{k} s_j^r(A) s_j^r(B), \quad 1 \le k \le n. \end{aligned}$$

Now use the statement II.3.5(vii). ∎

Corollary IV.2.6 *(Hölder's Inequality for Unitarily Invariant Norms) For every unitarily invariant norm and for all $A, B \in \mathbf{M(n)}$*

$$|||AB||| \leq |||\ |A|^p|||^{1/p}\ |||\ |A|^q|||^{1/q} \tag{IV.42}$$

for all $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. Use the special case of (IV.41) for $r = 1$ to get

$$\Phi\left(s(AB)\right) \leq \Phi\left(s(A)s(B)\right)$$

for every symmetric gauge function. Now use Theorem IV.1.6 and the fact that $(s(A))^p = s(|A|^p)$. ∎

Exercise IV.2.7 *Let p, q, r be positive real numbers with $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$. Then for every unitarily invariant norm*

$$|||\ |AB|^r|||^{1/r} \leq |||\ |A|^p|||^{1/p}|||\ |B|^q|||^{1/q}. \tag{IV.43}$$

Choosing $p = q = 1$, one gets from this

$$|||\ |AB|^{1/2}||| \leq (|||A|||\ |||B|||)^{1/2}. \tag{IV.44}$$

This is the **Cauchy-Schwarz inequality for unitarily invariant norms.**

Exercise IV.2.8 *Given a unitarily invariant norm $|||\cdot|||$ on $\mathbf{M(n)}$, define*

$$|||A|||^{(p)} = |||\ |A|^p|||^{1/p} \quad 1 \leq p < \infty. \tag{IV.45}$$

Show that this is a unitarily invariant norm. Note that

$$||A||_{p_1}^{(p_2)} = ||A||_{p_1 p_2} \quad \textit{for all} \quad p_1, p_2 \geq 1 \tag{IV.46}$$

and

$$||A||_{(k)}^{(p)} = (\sum_{j=1}^{k} s_j^p(A))^{1/p} \quad \textit{for} \quad p \geq 1, 1 \leq k \leq n. \tag{IV.47}$$

Definition IV.2.9 *A unitarily invariant norm on $\mathbf{M(n)}$ is called a* **Q-norm** *if it corresponds to a quadratic symmetric gauge function; i.e., $|||\cdot|||$ is a Q-norm if and only if there exists a unitarily invariant norm $|||\cdot|||^{\wedge}$ such that*

$$|||A|||^2 = |||A^*A|||^{\wedge}. \tag{IV.48}$$

Note that the norm $||\ ||_p$ is a Q-norm if and only if $p \geq 2$ because

$$||A||_p^2 = ||A^*A||_{p/2}. \tag{IV.49}$$

The norms defined in (IV.47) are Q-norms if and only if $p \geq 2$.

Exercise IV.2.10 *Let $\|\cdot\|_Q$ denote a Q-norm. Observe that the following conditions are equivalent:*

(i) $\|A\|_Q \leq \|B\|_Q$ *for all Q-norms.*

(ii) $|||A^*A||| \leq |||B^*B|||$ *for all unitarily invariant norms.*

(iii) $\|A\|_{(k)}^{(2)} \leq \|B\|_{(k)}^{(2)}$ *for* $k = 1, 2, \ldots, n$.

(iv) $(s(A))^2 \prec_w (s(B))^2$.

Duality in the space of unitarily invariant norms is defined via the inner product (IV.35). If $|||\cdot|||$ is a unitarily invariant norm, define $|||\cdot|||'$ as

$$|||A|||' = \sup_{|||B|||=1} |\langle A, B\rangle| = \sup_{|||B|||=1} |\mathrm{tr} A^*B|. \tag{IV.50}$$

It is easy to see that this defines a norm that is unitarily invariant.

Proposition IV.2.11 *Let Φ be a symmetric gauge function on $\mathbb{R}^n$ and let $\|\cdot\|_\Phi$ be the corresponding unitarily invariant norm on $\mathbf{M(n)}$. Then $\|\cdot\|'_\Phi = \|\cdot\|_{\Phi'}$.*

Proof. We have from (II.40) and (IV.41)

$$|\mathrm{tr} A^*B| \leq \mathrm{tr}|A^*B| = \sum_{j=1}^{n} s_j(A^*B) \leq \sum_{j=1}^{n} s_j(A)s_j(B).$$

It follows that

$$\|A\|'_\Phi \leq \Phi'(s(A)) = \|A\|_{\Phi'}.$$

Conversely,

$$\begin{aligned}
\|A\|_{\Phi'} &= \Phi'(s(A)) \\
&= \sup\left\{\sum_{j=1}^{n} s_j(A)y_j : y \in \mathbb{R}^n, \Phi(y) = 1\right\} \\
&= \sup\left\{\mathrm{tr}[\mathrm{diag}(s(A))\mathrm{diag}(y)] : \|\mathrm{diag}(y)\|_\Phi = 1\right\} \\
&\leq \|\mathrm{diag}(s(A))\|'_\Phi = \|A\|'_\Phi.
\end{aligned}$$

■

Exercise IV.2.12 *From statements about duals proved in Section IV.1, we can now conclude that*

(i) $|\mathrm{tr}\, A^*B| \leq |||A||| \cdot |||B|||'$ *for every unitarily invariant norm.*

(ii) $\|A\|'_p = \|A\|_q$ *for* $1 \leq p \leq \infty, \frac{1}{p} + \frac{1}{q} = 1$.

(iii) $\|A\|'_{(k)} = \max\{\|A\|_{(1)}, \frac{1}{k}\|A\|_{(n)}\},\ 1 \leq k \leq n$.

(iv) *The only unitarily invariant norm that is its own dual is the Hilbert-Schmidt norm* $\|\cdot\|_2$.

(v) *The only norm that is a Q-norm and is also the dual of a Q-norm is the norm* $\|\cdot\|_2$.

Duals of Q-norms will be called Q'-norms. These include the norms $\|\cdot\|_p, 1 \le p \le 2$.

An important property of all unitarily invariant norms is that they are all reduced by **pinchings**. If $P_1, \ldots, P_k$ are mutually orthogonal projections such that $P_1 \oplus P_2 \oplus \ldots \oplus P_k = I$, then the operator on $\mathbf{M(n)}$ defined as

$$\mathcal{C}(A) = \sum_{j=1}^{k} P_j A P_j \tag{IV.51}$$

is called a pinching operator. It is easy to see that

$$|||\mathcal{C}(A)||| \le |||A||| \tag{IV.52}$$

for every unitarily invariant norm. (See Problem II.5.5.) We will call this the **pinching inequality.**

Let us illustrate one use of this inequality.

Theorem IV.2.13 *Let* $A, B \in \mathbf{M(n)}$. *Then for every unitarily invariant norm on* $\mathbf{M(2n)}$

$$\frac{1}{2}\left|\left|\left|\begin{bmatrix} A+B & 0 \\ 0 & A+B \end{bmatrix}\right|\right|\right| \le \left|\left|\left|\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right|\right|\right| \le \left|\left|\left|\begin{bmatrix} |A|+|B| & 0 \\ 0 & 0 \end{bmatrix}\right|\right|\right|. \tag{IV.53}$$

Proof. The first inequality follows easily from the observation that $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ and $\begin{bmatrix} B & 0 \\ 0 & A \end{bmatrix}$ are unitarily equivalent.

If we prove the second inequality in the special case when A, B are positive, the general case follows easily. So, assume A, B are positive. Then

$$\begin{bmatrix} A+B & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} A^{1/2} & B^{1/2} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A^{1/2} & 0 \\ B^{1/2} & 0 \end{bmatrix},$$

where $A^{1/2}, B^{1/2}$ are the positive square roots of A, B. Since T^*T and TT^* are unitarily equivalent for every T, the matrix $\begin{bmatrix} A+B & 0 \\ 0 & 0 \end{bmatrix}$ is unitarily equivalent to

$$\begin{bmatrix} A^{1/2} & 0 \\ B^{1/2} & 0 \end{bmatrix}\begin{bmatrix} A^{1/2} & B^{1/2} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} A & A^{1/2}B^{1/2} \\ B^{1/2}A^{1/2} & B \end{bmatrix}.$$

But $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is a pinching of this last matrix. ■

As a corollary we have:

Theorem IV.2.14 *(Rotfel'd) Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function with $f(0) = 0$. Then the function F on $\mathbf{M(n)}$ defined by*

$$F(A) = \sum_{j=1}^{n} f(s_j(A)) \tag{IV.54}$$

is subadditive.

Proof. The second inequality in (IV.53) can be written as a majorisation in $\mathbb{R}^{2n}$:

$$(s(A), s(B)) \prec_w (s(|A| + |B|), 0)$$

for all $A, B \in \mathbf{M(n)}$. We also know that $s(|A| + |B|) \prec s(A) + s(B)$. Hence

$$(s(A), s(B)) \prec (s(A) + s(B), 0)\,.$$

Now proceed as in Problem II.5.12. ■

Exercise IV.2.15 *Let $|||\cdot|||$ be a unitarily invariant norm on $\mathbf{M(n)}$. For $m < n$ and $A \in \mathbf{M(m)}$, define*

$$|||A|||^\dagger = \left|\left|\left| \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \right|\right|\right|.$$

Show that $|||\cdot|||^\dagger$ defines a unitarily invariant norm on $\mathbf{M(m)}$.

We will use this idea of "dilating" A and of going from $\mathbf{M(n)}$ to $\mathbf{M(2n)}$ in later chapters. Procedures given in Exercises IV.1.19 and IV.1.20 can be adapted to matrices to generate unitarily invariant norms.

IV.3 Lidskii's Theorem (Third Proof)

Let $\lambda^\downarrow(A)$ denote the n-vector whose coordinates are the eigenvalues of a Hermitian matrix A arranged in decreasing order. Lidskii's Theorem, for which we gave two proofs in Section III.4, says that if A, B are Hermitian matrices, then we have the majorisation

$$\lambda^\downarrow(A) - \lambda^\downarrow(B) \prec \lambda(A - B). \tag{IV.55}$$

We will give another proof of this theorem now, using the easier ideas of Sections III.1 and III.2.

Exercise IV.3.1 *One corollary of Lidskii's Theorem is that, if A and B are any two matrices, then*

$$|s(A) - s(B)| \prec_w s(A - B). \tag{IV.56}$$

See Problem III.6.13. Conversely, show that if (IV.56) is known to be true for all matrices A, B, then we can derive from it the statement (IV.55). [Hint: Choose real numbers α, β such that $A + \alpha I \geq B + \beta I \geq 0$.]

We will prove (IV.56) by a different argument. To prove this we need to prove that for each of the Ky Fan symmetric gauge functions $\Phi_{(k)}, 1 \leq k \leq n$, we have the inequality

$$\Phi_{(k)}\left(s(A) - s(B)\right) \leq \Phi_{(k)}\left(s(A - B)\right). \tag{IV.57}$$

We will prove this for $\Phi_{(1)}$ and $\Phi_{(n)}$, and then use the interpolation formulas (IV.9) and (IV.39).

For $\Phi_{(1)}$ this is easy. By Weyl's perturbation theorem (Corollary III.2.6) we have

$$\max_j |\lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B)| \leq ||A - B||.$$

This can be proved easily by another argument also. For any j consider the subspaces spanned by $\{u_1, \ldots, u_j\}$ and $\{v_j, \ldots, v_n\}$, where $u_i, v_i, 1 \leq i \leq n$ are eigenvectors of A and B corresponding to their eigenvalues $\lambda_i^{\downarrow}(A)$ and $\lambda_i^{\downarrow}(B)$, respectively. Since the dimensions of these two spaces add up to $n + 1$, they have a nonzero intersection. For a unit vector x in this intersection we have $\langle x, Ax \rangle \geq \lambda_j^{\downarrow}(A)$ and $\langle x, Bx \rangle \leq \lambda_j^{\downarrow}(B)$. Hence, we have

$$||A - B|| \geq |\langle x, (A - B)x \rangle| \geq \lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B).$$

So, by symmetry

$$|\lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B)| \leq ||A - B||, \quad 1 \leq j \leq n.$$

From this, as before, we can get

$$\max_j |s_j(A) - s_j(B)| \leq ||A - B||$$

for any two matrices A and B. This is the same as saying

$$\Phi_{(1)}\left(s(A) - s(B)\right) \leq \Phi_{(1)}\left(s(A - B)\right). \tag{IV.58}$$

Let T be a Hermitian matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > \lambda_{p+1} \geq \cdots \geq \lambda_n$, where $\lambda_p \geq 0 > \lambda_{p+1}$. Choose a unitary matrix U such that $T = UDU^*$, where D is the diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Let $D^+ = (\lambda_1, \ldots, \lambda_p, 0, \cdots, 0)$ and $D^- = (0, \cdots, 0, -\lambda_{p+1}, \ldots, -\lambda_n)$. Let $T^+ = UD^+U^*$, $T^- = UD^-U^*$. Then both T^+ and T^- are positive matrices and

$$T = T^+ - T^-. \tag{IV.59}$$

This is called the **Jordan decomposition** of T.

Lemma IV.3.2 *If A, B are $n \times n$ Hermitian matrices, then*

$$\sum_{j=1}^{n} |\lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B)| \leq ||A - B||_{(n)}. \qquad \text{(IV.60)}$$

Proof. Using the Jordan decomposition of $A - B$ we can write

$$||A - B||_{(n)} = \operatorname{tr}(A - B)^+ + \operatorname{tr}(A - B)^-.$$

If we put

$$C = A + (A - B)^- = B + (A - B)^+,$$

then $C \geq A$ and $C \geq B$. Hence, by Weyl's monotonicity principle, $\lambda_j^{\downarrow}(C) \geq \lambda_j^{\downarrow}(A)$ and $\lambda_j^{\downarrow}(C) \geq \lambda_j^{\downarrow}(B)$ for all j. From these inequalities it follows that

$$|\lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B)| \leq \lambda_j^{\downarrow}(2C) - \lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B).$$

Hence,

$$\sum_{j=1}^{n} |\lambda_j^{\downarrow}(A) - \lambda_j^{\downarrow}(B)| \leq \operatorname{tr}(2C - A - B) = ||A - B||_{(n)}.$$

■

Corollary IV.3.3 *For any two $n \times n$ matrices A, B we have*

$$\Phi_{(n)}\,(s(A) - s(B)) = \sum_{j=1}^{n} |s_j(A) - s_j(B)| \leq ||A - B||_{(n)}. \qquad \text{(IV.61)}$$

Theorem IV.3.4 *For $n \times n$ matrices A, B we have the majorisation*

$$|s(A) - s(B)| \prec_w s(A - B).$$

Proof. Choose any index $k = 1, 2, \ldots, n$ and fix it. By Proposition IV.2.3, there exist $X, Y \in \mathbf{M(n)}$ such that

$$A - B = X + Y$$

and

$$||A - B||_{(k)} = ||X||_{(n)} + k||Y||.$$

Define vectors α, β as

$$\begin{aligned} \alpha &= s(X + B) - s(B), \\ \beta &= s(A) - s(X + B). \end{aligned}$$

Then

$$s(A) - s(B) = \alpha + \beta.$$

Hence, by Proposition IV.1.5 (or Proposition IV.2.3 restricted to diagonal matrices) and by (IV.58) and (IV.61), we have

$$\begin{aligned}\Phi_{(k)}(s(A)-s(B)) &\leq \Phi_{(n)}(\alpha)+k\,\Phi_{(1)}(\beta)\\ &= \Phi_{(n)}(s(X+B)-s(B))+k\,\Phi_{(1)}(s(A)-s(X+B))\\ &\leq ||X||_{(n)}+k||A-(X+B)||\\ &= ||X||_{(n)}+k||Y||\\ &= ||A-B||_{(k)}.\end{aligned}$$

This proves the theorem. ∎

As we observed in Exercise IV.3.1, this theorem is equivalent to Lidskii's Theorem.

In Section III.2 we introduced the notation Eig A for a diagonal matrix whose diagonal entries are the eigenvalues of a matrix A. The majorisations

$$\lambda^{\downarrow}(A)-\lambda^{\downarrow}(B)\prec\lambda(A-B)\prec\lambda^{\downarrow}(A)-\lambda^{\uparrow}(B)$$

for the eigenvalues of Hermitian matrices lead to norm inequalities

$$|||\mathrm{Eig}^{\downarrow}(A)-\mathrm{Eig}^{\downarrow}(B)|||\leq|||A-B|||\leq|||\mathrm{Eig}^{\downarrow}(A)-\mathrm{Eig}^{\uparrow}(B)|||, \quad \text{(IV.62)}$$

for all unitarily invariant norms. This is just another way of expressing Theorem III.4.4. The inequalities of Theorem III.2.8 and Problem III.6.15 are special cases of this.

We will see several generalisations of this inequality and still other proofs of it.

Exercise IV.3.5 *If* $\mathrm{Sing}^{\downarrow}(A)$ *denotes the diagonal matrix whose diagonal entries are* $s_1(A),\ldots,s_n(A)$, *then it follows from Theorem IV.3.4 that for any two matrices* A, B

$$|||\mathrm{Sing}^{\downarrow}(A)-\mathrm{Sing}^{\downarrow}(B)|||\leq|||A-B|||$$

for every unitarily invariant norm. Show that in this case the "opposite inequality"

$$|||A-B|||\leq|||\mathrm{Sing}^{\downarrow}(A)-\mathrm{Sing}^{\uparrow}(B)|||$$

is not always true.

IV.4 Weakly Unitarily Invariant Norms

Consider the following numbers associated with an $n\times n$ matrix:

(i) $|\mathrm{tr}\,A|=|\sum\lambda_j(A)|$;

(ii) $\operatorname{spr} A = \max_{1 \le j \le n} |\lambda_j(A)|$, the **spectral radius** of A;

(iii) $w(A) = \max_{||x||=1} |\langle x, Ax \rangle|$, the **numerical radius** of A.

Of these, the first one is a seminorm but not a norm on $\mathbf{M}(\mathbf{n})$, the second one is not a seminorm, and the third one is a norm. (See Exercise I.2.10.)

All three functions of a matrix described above have an important invariance property: they do not change under **unitary conjugations**; i.e., the transformations $A \to UAU^*, U$ unitary, do not change these functions. Indeed, the first two are invariant under the larger class of **similarity transformations** $A \to SAS^{-1}, S$ invertible. The third one is not invariant under all such transformations.

Exercise IV.4.1 *Show that no norm on* $\mathbf{M}(\mathbf{n})$ *can be invariant under all similarity transformations.*

Unlike the norms that were studied in Section 2, none of the three functions mentioned above is invariant under all transformations $A \to UAV$, where U, V vary over the unitary group $\mathbf{U}(\mathbf{n})$.

We will call a norm τ on $\mathbf{M}(\mathbf{n})$ **weakly unitarily invariant** (**wui**, for short) if

$$\tau(A) = \tau(UAU^*) \quad \text{for all} \quad A \in \mathbf{M}(\mathbf{n}), U \in \mathbf{U}(\mathbf{n}). \tag{IV.63}$$

Examples of such norms include the unitarily invariant norms and the numerical radius. Some more will be constructed now.

Exercise IV.4.2 *Let* E_{11} *be the diagonal matrix with its top left entry 1 and all other entries zero. Then*

$$w(A) = \max_{U \in \mathbf{U}(\mathbf{n})} |\operatorname{tr} E_{11} UAU^*|. \tag{IV.64}$$

Equivalently,

$$w(A) = \max\{|\operatorname{tr} AP| : P \textit{ is an orthogonal projection of rank 1}\}.$$

Given a matrix C, let

$$w_C(A) = \max_{U \in \mathbf{U}(\mathbf{n})} |\operatorname{tr} CUAU^*|, \quad A \in \mathbf{M}(\mathbf{n}). \tag{IV.65}$$

This is called the **C-numerical radius** of A.

Exercise IV.4.3 *For every* $C \in \mathbf{M}(\mathbf{n})$, *the C-numerical radius* w_C *is a wui seminorm on* $\mathbf{M}(\mathbf{n})$.

Proposition IV.4.4 *The C-numerical radius* w_C *is a norm on* $\mathbf{M}(\mathbf{n})$ *if and only if*

(i) C *is not a scalar multiple of* I, *and*

(ii) $\operatorname{tr} C \neq 0$.

Proof. If $C = \lambda I$ for any $\lambda \in \mathbb{C}$, then $w_C(A) = |\lambda|\,|\text{tr}\,A|$, and this is zero if $\text{tr}\,A = 0$. So w_C cannot be a norm. If $\text{tr}\,C = 0$, then $w_C(I) = |\text{tr}\,C| = 0$. Again w_C is not a norm. Thus (i) and (ii) are necessary conditions for w_C to be a norm.

Conversely, suppose $w_C(A) = 0$. If A were a scalar multiple of I, this would mean that $\text{tr}\,C = 0$. So, if $\text{tr}\,C \neq 0$, then A is not a scalar multiple of I. Hence A has an eigenspace $\mathcal{M}$ of dimension m, for some $0 < m < n$. Since e^{tK} is a unitary matrix for all real t and skew-Hermitian K, the condition $w_C(A) = 0$ implies in particular that

$$\text{tr}\,C e^{tK} A e^{-tK} = 0 \quad \text{if} \quad t \in \mathbb{R}, K = -K^*.$$

Differentiating this relation at $t = 0$, one gets

$$\text{tr}\,(AC - CA)K = 0 \quad \text{if} \quad K = -K^*.$$

Hence, we also have

$$\text{tr}\,(AC - CA)X = 0 \ \text{ for all } \ X \in \mathbf{M(n)}.$$

Hence $AC = CA$. (Recall that $\langle S, T\rangle = \text{tr}S^*T$ is an inner product on $\mathbf{M(n)}$.) Since C commutes with A, it leaves invariant the m-dimensional eigenspace $\mathcal{M}$ of A we mentioned earlier. Now, note that since $w_C(A) = w_C(UAU^*), C$ also commutes with UAU^* for every $U \in \mathbf{U(n)}$. But UAU^* has the space $U\mathcal{M}$ as an eigenspace. So, C also leaves $U\mathcal{M}$ invariant for all $U \in \mathbf{U(n)}$. But this would mean that C leaves all m-dimensional subspaces invariant, which in turn would mean C leaves all one-dimensional subspaces invariant, which is possible only if C is a scalar multiple of I. ∎

More examples of wui norms are given in the following exercise.

Exercise IV.4.5 *(i)* $\tau(A) = ||A|| + |\text{tr}\,A|$ *is a wui norm. More generally, the sum of any wui norm and a wui seminorm is a wui norm.*

(ii) $\tau(A) = \max(||A||, |\text{tr}\,A|)$ *is a wui norm. More generally, the maximum of any wui norm and a wui seminorm is a wui norm.*

(iii) Let $W(A)$ *be the numerical range of* A*. Then its diameter* $\text{diam}\,W(A)$ *is a wui seminorm on* $\mathbf{M(n)}$*. It can be used to generate wui norms as in (i) and (ii). Of particular interest would be the norm* $\tau(A) = w(A) + \text{diam}\,W(A)$*.*

(iv) Let $m(A)$ *be any norm on* $\mathbf{M(n)}$*. Then*

$$\tau(A) = \max_{U \in \mathbf{U(n)}} m(UAU^*)$$

is a wui norm.

(v) Let $m(A)$ be any norm on $\mathbf{M(n)}$. Then

$$\tau(A) = \int_{\mathbf{U(n)}} m(UAU^*)dU,$$

where the integral is with respect to the (normalised) Haar measure on $\mathbf{U(n)}$ is a wui norm.

(vi) Let

$$\tau(A) = \max_{e_1,\ldots,e_n} \max_{i,j} |\langle e_i, Ae_j\rangle|,$$

where $e_1, \ldots, e_n$ varies over all orthonormal bases. Then τ is a wui norm. How is this related to (ii) and (iv) above?

Let S be the unit sphere in $\mathbb{C}^n$,

$$S = \{x \in \mathbb{C}^n : ||x|| = 1\},$$

and let $C(S)$ be the space of all complex valued continuous functions on S. Let dx denote the normalised Lebesgue measure on S. Consider the familiar L_p-norms on $C(S)$ defined as

$$\begin{aligned} N_p(f) &= ||f||_p = (\int_S |f(x)|^p dx)^{1/p}, \quad 1 \leq p < \infty, \\ N_\infty(f) &= ||f||_\infty = \max_{x \in S} |f(x)|. \end{aligned} \tag{IV.66}$$

Since the measure dx is invariant under rotations, the above norms satisfy the invariance property

$$N_p(f \circ U) = N_p(f) \ \text{ for all } f \in C(S), U \in \mathbf{U(n)}.$$

We will call a norm N on $C(S)$ a **unitarily invariant function norm** if

$$N(f \circ U) = N(f) \ \text{ for all } f \in C(S), U \in \mathbf{U(n)}. \tag{IV.67}$$

The L_p-norms are important examples of such norms.

Now, every $A \in \mathbf{M(n)}$ induces, naturally, a function f_A on S by its quadratic form:

$$f_A(x) = \langle x, Ax\rangle. \tag{IV.68}$$

The correspondence $A \to f_A$ is a linear map from $\mathbf{M(n)}$ into $C(S)$, which is one-to-one. So, given a norm N on $C(S)$, if we define a function N' on $\mathbf{M(n)}$ as

$$N'(A) = N(f_A), \tag{IV.69}$$

then N' is a norm on $\mathbf{M(n)}$. Further,

$$N'(UAU^*) = N(f_{UAU^*}) = N(f_A \circ U^*).$$

So, if N is a unitarily invariant function norm on $C(S)$ then N' is a wui norm on $\mathbf{M(n)}$. The next theorem says that all wui norms arise in this way:

Theorem IV.4.6 *A norm τ on $\mathbf{M(n)}$ is weakly unitarily invariant if and only if there exists a unitarily invariant function norm N on $C(S)$ such that $\tau = N'$, where the map $N \to N'$ is defined by relations (IV.68) and (IV.69).*

Proof. We need to prove that every wui norm τ on $\mathbf{M(n)}$ is of the form N' for some unitarily invariant function norm N.

Let $F = \{f_A : A \in \mathbf{M(n)}\}$. This is a finite-dimensional linear subspace of $C(S)$. Given a wui norm τ, define N_0 on F by

$$N_0(f_A) = \tau(A). \tag{IV.70}$$

Then N_0 defines a norm on F, and further, $N_0(f \circ U) = N_0(f)$ for all $f \in F$. We will extend N_0 from F to all of $C(S)$ to obtain a norm N that is unitarily invariant. Clearly, then $\tau = N'$.

This extension is obtained by an application of the Hahn-Banach Theorem. The space $C(S)$ is a Banach space with the supremum norm $||f||_\infty$. The finite-dimensional subspace F has two norms N_0 and $||\cdot||_\infty$. These must be equivalent: there exist constants $0 < \alpha \leq \beta < \infty$ such that $\alpha||f||_\infty \leq N_0(f) \leq \beta||f||_\infty$ for all $f \in F$. Let G be the set of all linear functionals on F that have norm less than or equal to 1 with respect to the norm N_0; i.e., the linear functional g is in G if and only if $|g(f)| \leq N_0(f)$ for all $f \in F$. By duality then $N_0(f) = \sup\limits_{g \in G}|g(f)|$, for every $f \in F$. Now $|g(f)| \leq \beta||f||_\infty$ for $g \in G$ and $f \in F$. Hence, by the Hahn-Banach Theorem, each g can be extended to a linear functional $\tilde{g}$ on $C(S)$ such that $|\tilde{g}(f)| \leq \beta||f||_\infty$ for all $f \in C(S)$. Now define

$$\theta(f) = \sup_{g \in G}|\tilde{g}(f)|, \quad \text{for all} \quad f \in C(S).$$

Then θ is a seminorm on $C(S)$ that coincides with N_0 on F. Let

$$\mu(f) = \max\{\theta(f), \alpha||f||_\infty\}, \qquad f \in C(S).$$

Then μ is a norm on $C(S)$, and μ coincides with N_0 on F. Now define

$$N(f) = \sup_{U \in \mathbf{U(n)}} \mu(f \circ U), \quad f \in C(S).$$

Then N is a unitarily invariant function norm on $C(S)$ that coincides with N_0 on F. The proof is complete. ∎

When $N = ||\cdot||_\infty$ the norm N' induced by the above procedure is the numerical radius w. Another example is discussed in the Notes.

The C-numerical radii play a useful role in proving inequalities for wui norms:

Theorem IV.4.7 *For $A, B \in \mathbf{M(n)}$ the following statements are equivalent:*

(i) $\tau(A) \leq \tau(B)$ *for all wui norms* τ.

(ii) $w_C(A) \leq w_C(B)$ *for all upper triangular matrices* C *that are not scalars and have nonzero trace.*

(iii) $w_C(A) \leq w_C(B)$ *for all* $C \in \mathbf{M(n)}$.

(iv) A *can be expressed as a finite sum* $A = \sum z_k U_k B U_k^*$ *where* $U_k \in \mathbf{U(n)}$ *and* z_k *are complex numbers with* $\sum |z_k| \leq 1$.

Proof. By Proposition IV.4.4, when C is not a scalar and $\operatorname{tr} C \neq 0$, each w_C is a wui norm. So (i) $\Rightarrow$ (ii).

Note that $w_C(A) = w_A(C)$ for all pairs of matrices A, C. So, if (ii) is true, then $w_A(C) \leq w_B(C)$ for all upper triangular nonscalar matrices C with nonzero trace. Since w_A and w_B are wui, and since every matrix is unitarily equivalent to an upper triangular matrix, this implies that $w_A(C) \leq w_B(C)$ for all nonscalar matrices C with nonzero trace. But such C are dense in the space $\mathbf{M(n)}$. So $w_A(C) \leq w_B(C)$ for all $C \in \mathbf{M(n)}$. Hence (iii) is true.

Let $\mathcal{K}$ be the convex hull of all matrices $e^{i\theta}UBU^*, \theta \in \mathbb{R}, U \in \mathbf{U(n)}$. Then $\mathcal{K}$ is a compact convex set in $\mathbf{M(n)}$. The statement (iv) is equivalent to saying that $A \in \mathcal{K}$. If $A \notin \mathcal{K}$, then by the Separating Hyperplane Theorem there exists a linear functional f on $\mathbf{M(n)}$ such that $\operatorname{Re} f(A) > \operatorname{Re} f(X)$ for all $X \in \mathcal{K}$. For this linear functional f there exists a matrix C such that $f(Y) = \operatorname{tr} CY$ for all $Y \in \mathbf{M(n)}$. (Problem IV.5.8) For these f and C we have

$$\begin{aligned} w_C(A) &= \max_{U \in \mathbf{U(n)}} |\operatorname{tr} CUAU^*| \geq |\operatorname{tr} CA| = |f(A)| \geq \operatorname{Re} f(A) \\ &> \max_{X \in \mathcal{K}} \operatorname{Re} f(X) \\ &= \max_{\theta, U} \operatorname{Re} \operatorname{tr} Ce^{i\theta}UBU^* \\ &= \max_{U} |\operatorname{tr} CUBU^*| \\ &= w_C(B). \end{aligned}$$

So, if (iii) were true, then (iv) cannot be false.

Clearly (iv) $\Rightarrow$ (i). ■

The family w_C of C-numerical radii, where C is not a scalar and has nonzero trace, thus plays a role analogous to that of the Ky Fan norms in the family of unitarily invariant norms. However, unlike the Ky Fan family on $\mathbf{M(n)}$, this family is infinite. It turns out that no finite subfamily of wui norms can play this role.

More precisely, there does not exist any finite family $\tau_1, \ldots, \tau_m$ of wui norms on $\mathbf{M(n)}$ that would lead to the inequalities $\tau(A) \leq \tau(B)$ for all wui norms whenever $\tau_j(A) \leq \tau_j(B), 1 \leq j \leq m$. For if such a family existed, then we would have

$$\{X : \tau(X) \leq \tau(I) \text{ for all wui norms } \tau\} = \bigcap_{j=1}^{m} \{X : \tau_j(X) \leq \tau_j(I)\}. \tag{IV.71}$$

Now each of the sets in this intersection contains 0 as an interior point (with respect to some fixed topology on $\mathbf{M(n)}$). Hence the intersection also contains 0 as an interior point. However, by Theorem IV.4.7, the set on the left-hand side of (IV.71) reduces to the set $\{zI : z \in \mathbb{C}, |z| \leq 1\}$, and this set has an empty interior in $\mathbf{M(n)}$.

Finally, note an important property of all wui norms:

$$\tau(\mathcal{C}(A)) \leq \tau(A) \tag{IV.72}$$

for all $A \in \mathbf{M(n)}$ and all pinchings $\mathcal{C}$ on $\mathbf{M(n)}$.

In Chapter 6 we will prove a generalisation of Lidskii's inequality (IV.62) extending it to all wui norms.

IV.5 Problems

Problem IV.5.1. When $0 < p < 1$, the function $\Phi_p(x) = (\sum |x_i|^p)^{1/p}$ does not define a norm. Show that in lieu of the triangle inequality we have

$$\Phi_p(x+y) \leq 2^{\frac{1}{p}-1}[\Phi_p(x) + \Phi_p(y)], \quad 0 < p < 1.$$

(Use the fact that $f(t) = t^p$ on $\mathbb{R}_+$ is subadditive when $0 < p \leq 1$ and convex when $p \geq 1$.)

Positive homogeneous functions that do not satisfy the triangle inequality but a weaker inequality $\varphi(x+y) \leq c[\varphi(x) + \varphi(y)]$ for some constant $c > 1$ are sometimes called **quasi-norms.**

Problem IV.5.2. More generally, show that for any symmetric gauge function Φ and $0 < p < 1$, if we define $\Phi^{(p)}$ as in (IV.17), then

$$\Phi^{(p)}(x+y) \leq 2^{\frac{1}{p}-1}[\Phi^{(p)}(x) + \Phi^{(p)}(y)], \quad 0 < p < 1.$$

Problem IV.5.3. All norms on $\mathbb{C}^n$ are equivalent in the sense that if Φ and Ψ are two norms, then there exists a constant K such that $\Phi(x) \leq K\Psi(x)$ for all $x \in \mathbb{C}^n$. Let

$$K_{\Phi,\Psi} = \inf\{K : \Phi(x) \leq K\Psi(x) \text{ for all } x\}.$$

Find the constants $K_{\Phi,\Psi}$ when Φ, Ψ are both members of the family Φ_p.

Problem IV.5.4. Show that for every norm Φ on $\mathbb{C}^n$ we have $\Phi'' = \Phi$; i.e., the dual of the dual of a norm is the norm itself.

Problem IV.5.5. Find the duals of the norms $\Phi_{(k)}^{(p)}$ defined by (IV.19). (These are somewhat complicated.)

Problem IV.5.6. For $0 < p < 1$ and a unitarily invariant norm $|||\cdot|||$ on $\mathbf{M(n)}$, let

$$|||A|||^{(p)} = |||\,|A|^p|||^{1/p}.$$

Show that

$$|||A+B|||^{(p)} \leq 2^{\frac{1}{p}-1}\left[|||A|||^{(p)} + |||B|||^{(p)}\right].$$

Problem IV.5.7. Choosing $p = q = 2$ in (IV.43) or (IV.42), one obtains

$$|||AB||| \leq |||A^*A|||^{1/2}\,|||B^*B|||^{1/2}.$$

This, like the inequality (IV.44), is also a form of the Cauchy-Schwarz inequality, for unitarily invariant norms. Show that this is just the inequality (IV.44) restricted to Q-norms.

Problem IV.5.8. Let f be any linear functional on $\mathbf{M(n)}$. Show that there exists a unique matrix X such that $f(A) = \operatorname{tr} XA$ for all $A \in \mathbf{M(n)}$.

Problem IV.5.9. Use Theorem IV.2.14 to show that for all $A, B \in \mathbf{M(n)}$

$$\det(1 + |A+B|) \leq \det(1+|A|)\det(1+|B|).$$

Problem IV.5.10. More generally, show that for $0 < p \leq 1$ and $\mu \geq 0$

$$\det(1 + \mu|A+B|^p) \leq \det(1+\mu|A|^p)\det(1+\mu|B|^p).$$

Problem IV.5.11. Let ℓ_p denote the space $\mathbb{C}^n$ with the p-norm defined in (IV.1) and (IV.2), $1 \leq p \leq \infty$. For a matrix A let $||A||_{p\to p'}$ denote the norm of A as a linear operator from ℓ_p to $\ell_{p'}$; i.e.,

$$||A||_{p\to p'} = \max_{||x||_p = 1} ||Ax||_{p'}.$$

Show that

$$\begin{aligned} ||A||_{1\to 1} &= \max_j \sum_i |a_{ij}|. \\ ||A||_{\infty\to\infty} &= \max_i \sum_j |a_{ij}|. \\ ||A||_{1\to\infty} &= \max_{i,j} |a_{ij}|. \end{aligned}$$

None of these norms is weakly unitarily invariant.

Problem IV.5.12. Show that there exists a weakly unitarily invariant norm τ such that $\tau(A) \neq \tau(A^*)$ for some $A \in \mathbf{M(n)}$.

Problem IV.5.13. Show that there exists a weakly unitarily invariant norm τ such that $\tau(A) > \tau(B)$ for some positive matrices A, B with $A \leq B$.

Problem IV.5.14. Let τ be a wui norm on $\mathbf{M(n)}$. Define ν on $\mathbf{M(n)}$ as $\nu(A) = \tau(|A|)$. Then ν is a unitarily invariant norm if and only if $\tau(A) \leq \tau(B)$ whenever $0 \leq A \leq B$.

Problem IV.5.15. Show that for every wui norm τ

$$\tau(\text{Eig } A) = \inf\{\tau(SAS^{-1}) : S \in \mathbf{GL(n)}\}.$$

When is the infimum attained?

Problem IV.5.16. Let τ be a wui norm on $\mathbf{M(n)}$. Show that for every A

$$\tau(A) \geq \frac{|\text{tr } A|}{n}\tau(I).$$

Use this to show that

$$\min\{\tau(A - B) : \text{tr } B = 0\} = \frac{|\text{tr } A|}{n}\tau(I).$$

IV.6 Notes and References

The first major paper on the theory of unitarily invariant norms and symmetric gauge functions was by J. von Neumann, *Some matrix inequalities and metrization of matric space,* Tomsk. Univ. Rev., 1(1937) 286-300, reprinted in his *Collected Works*, Pergamon Press, 1962. A famous book devoted to the study of such norms (for compact operators in a Hilbert space) is R. Schatten, *Norm Ideals of Completely Continuous Operators*, Springer-Verlag, 1960. Other excellent sources of information are the books by I.C. Gohberg and M.G. Krein cited in Chapter III, by A. Marshall and I. Olkin cited in Chapter II, and by R. Horn and C.R. Johnson cited in Chapter I. A succinct but complete summary can be found in L. Mirsky's paper cited in Chapter III. Much more on matrix norms (not necessarily invariant ones) can be found in the book *Matrix Norms and Their Applications*, by G.R. Belitskii and Y.I. Lyubich, Birkhäuser, 1988.

The notion of Q-norms is mentioned explicitly in R. Bhatia, *Some inequalities for norm ideals*, Commun. Math. Phys., 111(1987) 33-39. (The possible usefulness of the idea was suggested by C. Davis.) The Cauchy-Schwarz inequality (IV.44) is proved in R. Bhatia, *Perturbation inequalities*

for the absolute value map in norm ideals of operators, J. Operator Theory, 19 (1988) 129-136. This, and a whole family of inequalities including the one in Problem IV.5.7, are studied in detail by R.A. Horn and R. Mathias in two papers, *An analog of the Cauchy-Schwarz inequality for Hadamard products and unitarily invariant norms*, SIAM J. Matrix Anal. Appl., 11 (1990) 481-498, and *Cauchy-Schwarz inequalities associated with positive semidefinite matrices*, Linear Algebra Appl., 142(1990) 63-82. Many of the other inequalities in this section occur in K. Okubo, *Hölder-type norm inequalities for Schur products of matrices*, Linear Algebra Appl., 91(1987) 13-28. A general study of these and related inequalities is made in R. Bhatia and C. Davis, *Relations of linking and duality between symmetric gauge functions*, Operator Theory: Advances and Applications, 73(1994) 127-137.

Theorems IV.2.13 and IV.2.14 were proved by S. Ju. Rotfel'd, *The singular values of a sum of completely continuous operators*, in Topics in Mathematical Physics, Consultants Bureau, 1969, Vol. 3, pp. 73-78. See also, R.C. Thompson, *Convex and concave functions of singular values of matrix sums*, Pacific J. Math., 66(1976) 285-290. The results of Problems IV.5.9 and IV.5.10 are also due to Rotfel'd.

The proof of Lidskii's Theorem given in Section IV.3 is adapted from F. Hiai and Y. Nakamura, *Majorisation for generalised s-numbers in semifinite von Neumann algebras*, Math. Z., 195(1987) 17-27.

The theory of weakly unitarily invariant norms was developed in R. Bhatia and J.A.R. Holbrook, *Unitary invariance and spectral variation*, Linear Algebra Appl., 95(1987) 43-68. Theorem IV.4.6 is proved in this paper. More on C-numerical radii can be found in C.-K. Li and N.-K. Tsing, *Norms that are invariant under unitary similarities and the C-numerical radii*, Linear and Multilinear Algebra, 24(1989) 209-222. Theorem IV.4.7 is taken from this paper. A part of this theorem (the equivalence of conditions (i) and (iv)) was proved in R. Bhatia and J.A.R. Holbrook, *A softer, stronger Lidskii theorem*, Proc. Indian Acad. Sciences (Math. Sciences), 99 (1989) 75-83. Two papers dealing with wui norms for infinite-dimensional operators are C.-K. Fong and J.A.R. Holbrook, *Unitarily invariant operator norms*, Canad. J. Math., 35 (1983) 274-299, and C.-K. Fong, H. Radjavi, and P. Rosenthal, *Norms for matrices and operators*, J. Operator Theory, 18(1987) 99-113.

The theory of wui norms is not developed as fully as that of unitarily invariant norms. Theorem IV.4.6 would be useful if one could make the correspondence between τ and N more explicit. As things stand, this has not been done even for some well-known and much-used norms like the L_p-norms. When N is the L_∞ function norm, we have noted that $N'(A) = w(A)$. When N is the L_2 function norm, then it is shown in the Bhatia-Holbrook (1987) paper cited above that

$$N'(A) = \left(\frac{||A||_2^2 + |\text{tr}\, A|^2}{n + n^2} \right)^{1/2}.$$

For other values of p, the correspondence has not been worked out.

For a recent survey of several results on invariant norms see C.-K. Li, *Some aspects of the theory of norms*, Linear Algebra Appl., 212/213 (1994) 71-100.

V

Operator Monotone and Operator Convex Functions

In this chapter we study an important and useful class of functions called operator monotone functions. These are real functions whose extensions to Hermitian matrices preserve order. Such functions have several special properties, some of which are studied in this chapter. They are closely related to properties of operator convex functions. We shall study both of these together.

V.1 Definitions and Simple Examples

Let f be a real function defined on an interval I. If $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ is a diagonal matrix whose diagonal entries λ_j are in I, we define $f(D) = \text{diag}(f(\lambda_1), \dots, f(\lambda_n))$. If A is a Hermitian matrix whose eigenvalues λ_j are in I, we choose a unitary U such that $A = UDU^*$, where D is diagonal, and then define $f(A) = Uf(D)U^*$. In this way we can define $f(A)$ for all Hermitian matrices (of any order) whose eigenvalues are in I. In the rest of this chapter, it will always be assumed that our functions are real functions defined on an interval (finite or infinite, closed or open) and are extended to Hermitian matrices in this way.

We will use the notation $A \le B$ to mean A and B are Hermitian and $B - A$ is positive. The relation $\le$ is a partial order on Hermitian matrices.

A function f is said to be **matrix monotone of order n** if it is monotone with respect to this order on $n \times n$ Hermitian matrices, i.e., if $A \le B$ implies $f(A) \le f(B)$. If f is matrix monotone of order n for all n we say f is **matrix monotone** or **operator monotone**.

A function f is said to be **matrix convex of order n** if for all $n \times n$ Hermitian matrices A and B and for all real numbers $0 \leq \lambda \leq 1$,

$$f((1-\lambda)A + \lambda B) \leq (1-\lambda)f(A) + \lambda f(B). \tag{V.1}$$

If f is matrix convex of all orders, we say that f is **matrix convex** or **operator convex.**

(Note that if the eigenvalues of A and B are all in an interval I, then the eigenvalues of any convex combination of A, B are also in I. This is an easy consequence of results in Chapter III.)

We will consider continuous functions only. In this case, the condition (V.1) can be replaced by the more special condition

$$f\left(\frac{A+B}{2}\right) \leq \frac{f(A)+f(B)}{2}. \tag{V.2}$$

(Functions satisfying (V.2) are called **mid-point operator convex**, and if they are continuous, then they are convex.)

A function f is called **operator concave** if the function $-f$ is operator convex.

It is clear that the set of operator monotone functions and the set of operator convex functions are both closed under positive linear combinations and also under (pointwise) limits. In other words, if f, g are operator monotone, and if α, β are positive real numbers, then $\alpha f + \beta g$ is also operator monotone. If f_n are operator monotone, and if $f_n(x) \rightarrow f(x)$, then f is also operator monotone. The same is true for operator convex functions.

Example V.1.1 *The function $f(t) = \alpha + \beta t$ is operator monotone (on every interval) for every $\alpha \in \mathbb{R}$ and $\beta \geq 0$. It is operator convex for all $\alpha, \beta \in \mathbb{R}$.*

The first surprise is in the following example.

Example V.1.2 *The function $f(t) = t^2$ on $[0, \infty)$ is not operator monotone. In other words, there exist positive matrices A, B such that $B - A$ is positive but $B^2 - A^2$ is not. To see this, take*

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

Example V.1.3 *The function $f(t) = t^2$ is operator convex on every interval. To see this, note that for any Hermitian matrices A, B,*

$$\frac{A^2+B^2}{2} - \left(\frac{A+B}{2}\right)^2 = \frac{1}{4}(A^2 + B^2 - AB - BA) = \frac{1}{4}(A-B)^2 \geq 0.$$

This shows that the function $f(t) = \alpha + \beta t + \gamma t^2$ is operator convex for all $\alpha, \beta \in \mathbb{R}$, $\gamma \geq 0$.

Example V.1.4 *The function $f(t) = t^3$ on $[0,\infty)$ is not operator convex. To see this, let*

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}.$$

Then,

$$\frac{A^3+B^3}{2} - \left(\frac{A+B}{2}\right)^3 = \begin{pmatrix} 6 & 1 \\ 1 & 0 \end{pmatrix},$$

and this is not positive.

Examples V.1.2 and V.1.4 show that very simple functions which are monotone (convex) as real functions need not be operator monotone (operator convex). A complete description of operator monotone and operator convex functions will be given in later sections. It is instructive to study a few more examples first. The operator monotonicity or convexity of some functions can be proved by special arguments that are useful in other contexts as well.

We will repeatedly use two simple facts. If A is positive, then $A \leq I$ if and only if $\operatorname{spr}(A) \leq 1$. An operator A is a contraction ($\|A\| \leq 1$) if and only if $A^*A \leq I$. This is also equivalent to the condition $AA^* \leq I$.

The following elementary lemma is also used often.

Lemma V.1.5 *If $B \geq A$, then for every operator X we have $X^*BX \geq X^*AX$.*

Proof. For every vector u we have,

$$\langle u, X^*BXu\rangle = \langle Xu, BXu\rangle \geq \langle Xu, AXu\rangle = \langle u, X^*AXu\rangle.$$

This proves the lemma.

An equally brief proof goes as follows. Let C be the positive square root of the positive operator $B - A$. Then

$$X^*(B-A)X = X^*CCX = (CX)^*CX \geq 0.$$

■

Proposition V.1.6 *The function $f(t) = -\frac{1}{t}$ is operator monotone on $(0,\infty)$.*

Proof. Let $B \geq A > 0$. Then, by Lemma V.1.5, $I \geq B^{-1/2}AB^{-1/2}$. Since the map $T \to T^{-1}$ is order-reversing on commuting positive operators, we have $I \leq B^{1/2}A^{-1}B^{1/2}$. Again, using Lemma V.1.5 we get from this $B^{-1} \leq A^{-1}$. ■

Lemma V.1.7 *If $B \geq A \geq 0$ and B is invertible, then $\|A^{1/2}B^{-1/2}\| \leq 1$.*

Proof. If $B \geq A \geq 0$, then $I \geq B^{-1/2}AB^{-1/2} = (A^{1/2}B^{-1/2})^*A^{1/2}B^{-1/2}$, and hence $\|A^{1/2}B^{-1/2}\| \leq 1$. ■

Proposition V.1.8 *The function $f(t) = t^{1/2}$ is operator monotone on $[0, \infty)$.*

Proof. Let $B \geq A \geq 0$. Suppose B is invertible. Then, by Lemma V.1.7,

$$1 \geq \|A^{1/2}B^{-1/2}\| \geq \text{spr}(A^{1/2}B^{-1/2}) = \text{spr}(B^{-1/4}A^{1/2}B^{-1/4}).$$

Since $B^{-1/4}AB^{-1/4}$ is positive, this implies that $I \geq B^{-1/4}A^{1/2}B^{-1/4}$. Hence, by Lemma V.1.5, $B^{1/2} \geq A^{1/2}$. This proves the proposition under the assumption that B is invertible. If B is not strictly positive, then for every $\varepsilon > 0, B + \varepsilon I$ is strictly positive. So, $(B + \varepsilon I)^{1/2} \geq A^{1/2}$. Let $\varepsilon \to 0$. This shows that $B^{1/2} \geq A^{1/2}$. ■

Theorem V.1.9 *The function $f(t) = t^r$ is operator monotone on $[0, \infty)$ for $0 \leq r \leq 1$.*

Proof. Let r be a dyadic rational, i.e., a number of the form $r = \frac{m}{2^n}$, where n is any positive integer and $1 \leq m \leq 2^n$. We will first prove the assertion for such r. This is done by induction on n.

Proposition V.1.8 shows that the assertion of the theorem is true when $n = 1$. Suppose it is also true for all dyadic rationals $\frac{m}{2^j}$, in which $1 \leq j \leq n-1$. Let $B \geq A$ and let $r = \frac{m}{2^n}$. Suppose $m \leq 2^{n-1}$. Then, by the induction hypothesis, $B^{m/2^{n-1}} \geq A^{m/2^{n-1}}$. Hence, by Proposition V.1.8, $B^{m/2^n} \geq A^{m/2^n}$. Suppose $m > 2^{n-1}$. If $B \geq A > 0$, then $A^{-1} \geq B^{-1}$. Using Lemma V.1.5, we have $B^{m/2^n} A^{-1} B^{m/2^n} \geq B^{m/2^n} B^{-1} B^{m/2^n} = B^{(m/2^{n-1}-1)}$. By the same argument,

$$\begin{aligned} A^{-1/2} B^{m/2^n} A^{-1} B^{m/2^n} A^{-1/2} &\geq A^{-1/2} B^{(m/2^{n-1}-1)} A^{-1/2} \\ &\geq A^{-1/2} A^{(m/2^{n-1}-1)} A^{-1/2} \end{aligned}$$

(by the induction hypothesis). This can be written also as

$$(A^{-1/2} B^{m/2^n} A^{-1/2})^2 \geq A^{(m/2^{n-1}-2)}.$$

So, by the operator monotonicity of the square root,

$$A^{-1/2} B^{m/2^n} A^{-1/2} \geq A^{(m/2^n-1)}.$$

Hence, $B^{m/2^n} \geq A^{m/2^n}$.

We have shown that $B \geq A > 0$ implies $B^r \geq A^r$ for all dyadic rationals r in $[0, 1]$. Such r are dense in $[0, 1]$. So we have $B^r \geq A^r$ for all r in $[0, 1]$. By continuity this is true even when A is positive semidefinite . ■

Exercise V.1.10 *Another proof of Theorem V.1.9 is outlined below. Fill in the details.*

(i) *The composition of two operator monotone functions is operator monotone. Use this and Proposition V.1.6 to prove that the function* $f(t) = \frac{t}{1+t}$ *is operator monotone on* $(0, \infty)$.

(ii) *For each* $\lambda > 0$, *the function* $f(t) = \frac{t}{\lambda+t}$ *is operator monotone on* $(0, \infty)$.

(iii) *One of the integrals calculated by contour integration in Complex Analysis is*

$$\int_0^\infty \frac{\lambda^{r-1}}{1+\lambda} d\lambda = \pi \operatorname{cosec} r\pi, \quad 0 < r < 1. \tag{V.3}$$

By a change of variables, obtain from this the formula

$$t^r = \frac{\sin r\pi}{\pi} \int_0^\infty \frac{t}{\lambda + t} \lambda^{r-1} d\lambda \tag{V.4}$$

valid for all $t > 0$ *and* $0 < r < 1$.

(iv) *Thus, we can write*

$$t^r = \int_0^\infty \frac{t}{\lambda + t}\, d\mu(\lambda), \quad 0 < r < 1, \tag{V.5}$$

where μ *is a positive measure on* $(0, \infty)$. *Now use (ii) to conclude that the function* $f(t) = t^r$ *is operator monotone on* $(0, \infty)$ *for* $0 \le r \le 1$.

Example V.1.11 *The function* $f(t) = |t|$ *is not operator convex on any interval that contains 0. To see this, take*

$$A = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Then

$$|A| = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad |A| + |B| = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}.$$

But $|A+B| = \sqrt{2}\, I$. *So* $|A|+|B|-|A+B|$ *is not positive. (See also Exercise III.5.7.)*

Example V.1.12 *The function* $f(t) = t \vee 0$ *is not operator convex on any interval that contains 0. To see this, take* A, B *as in Example V.1.11. Since the eigenvalues of* A *are* -2 *and 0,* $f(A) = 0$. *So* $\frac{1}{2}(f(A) + f(B)) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. *Any positive matrix dominated by this must have* $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ *as an eigenvector with 0 as the corresponding eigenvalue. Since* $\frac{1}{2}(A + B)$ *does not have* $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ *as an eigenvector, neither does* $f(\frac{A+B}{2})$.

Exercise V.1.13 *Let I be any interval. For $a \in I$, let $f(t) = (t-a) \vee 0$. Then f is called an "angle function" angled at a. If I is a finite interval, then every convex function on I is a limit of positive linear combinations of linear functions and angle functions. Use this to show that angle functions are not operator convex.*

Exercise V.1.14 *Show that the function $f(t) = t \vee 0$ is not operator monotone on any interval that contains 0.*

Exercise V.1.15 *Let A, B be positive. Show that*

$$\frac{A^{-1}+B^{-1}}{2} - \left(\frac{A+B}{2}\right)^{-1} = \frac{(A^{-1}-B^{-1})(A^{-1}+B^{-1})^{-1}(A^{-1}-B^{-1})}{2}.$$

Therefore, the function $f(t) = \frac{1}{t}$ is operator convex on $(0,\infty)$.

V.2 Some Characterisations

There are several different notions of averaging in the space of operators. In this section we study the relationship between some of these operations and operator convex functions. This leads to some characterisations of operator convex and operator monotone functions and to the interrelations between them.

In the proofs that are to follow, we will frequently use properties of operators on the direct sum $\mathcal{H} \oplus \mathcal{H}$ to draw conclusions about operators on $\mathcal{H}$. This technique was outlined briefly in Section I.3.

Let K be a contraction on $\mathcal{H}$. Let $L = (I-KK^*)^{1/2}$, $M = (I-K^*K)^{1/2}$. Then the operators U, V defined as

$$U = \begin{pmatrix} K & L \\ M & -K^* \end{pmatrix}, \quad V = \begin{pmatrix} K & -L \\ M & K^* \end{pmatrix} \tag{V.6}$$

are unitary operators on $\mathcal{H} \oplus \mathcal{H}$. (See Exercise I.3.6.) More specially, for each $0 \le \lambda \le 1$, the operator

$$W = \begin{pmatrix} \lambda^{1/2} I & -(1-\lambda)^{1/2} I \\ (1-\lambda)^{1/2} I & \lambda^{1/2} I \end{pmatrix} \tag{V.7}$$

is a unitary operator on $\mathcal{H} \oplus \mathcal{H}$.

Theorem V.2.1 *Let f be a real function on an interval I. Then the following two statements are equivalent:*

(i) f is operator convex on I.

(ii) $f(\mathcal{C}(A)) \le \mathcal{C}(f(A))$ for every Hermitian operator A (on a Hilbert space $\mathcal{H}$) whose spectrum is contained in I and for every pinching $\mathcal{C}$ (in the space $\mathcal{H}$).

Proof. (i) $\Rightarrow$(ii): Every pinching is a product of pinchings by two complementary projections. (See Problems II.5.4 and II.5.5.) So we need to prove this implication only for pinchings $\mathcal{C}$ of the form

$$\mathcal{C}(X) = \frac{X + U^*XU}{2}, \quad \text{where } U = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

For such a $\mathcal{C}$

$$\begin{aligned} f(\mathcal{C}(A)) &= f\left(\frac{A + U^*AU}{2}\right) \leq \frac{f(A) + f(U^*AU)}{2} \\ &= \frac{f(A) + U^*f(A)U}{2} = \mathcal{C}(f(A)). \end{aligned}$$

(ii) $\Rightarrow$ (i): Let A, B be Hermitian operators on $\mathcal{H}$, both having their spectrum in I. Consider the operator $T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ on $\mathcal{H} \oplus \mathcal{H}$. If W is the unitary operator defined in (V.7), then the diagonal entries of W^*TW are $\lambda A + (1-\lambda)B$ and $(1-\lambda)A + \lambda B$. So if $\mathcal{C}$ is the pinching on $\mathcal{H} \oplus \mathcal{H}$ induced by the projections onto the two summands, then

$$\mathcal{C}(W^*TW) = \begin{pmatrix} \lambda A + (1-\lambda)B & 0 \\ 0 & (1-\lambda)A + \lambda B \end{pmatrix}.$$

By the same argument,

$$\begin{aligned} \mathcal{C}(f(W^*TW)) &= \mathcal{C}(W^*f(T)W) \\ &= \begin{pmatrix} \lambda f(A) + (1-\lambda)f(B) & 0 \\ 0 & (1-\lambda)f(A) + \lambda f(B) \end{pmatrix}. \end{aligned}$$

So the condition $f(\mathcal{C}(W^*TW)) \leq \mathcal{C}(f(W^*TW))$ implies that

$$f(\lambda A + (1-\lambda)B) \leq \lambda f(A) + (1-\lambda)f(B).$$

■

Exercise V.2.2 *The following conditions are equivalent:*

(i) f *is operator convex on* I.

(ii) $f(A_{\mathcal{M}}) \leq (f(A))_{\mathcal{M}}$ *for every Hermitian operator* A *with its spectrum in* I, *and for every compression* $T \to T_{\mathcal{M}}$.

(iii) $f(V^*AV) \leq V^*f(A)V$ *for every Hermitian operator* A *(on* $\mathcal{H}$*) with its spectrum in* I, *and for every isometry from any Hilbert space into* $\mathcal{H}$.

(See Section III.1 for the definition of a compression.)

Theorem V.2.3 *Let I be an interval containing 0 and let f be a real function on I. Then the following conditions are equivalent:*

(i) f *is operator convex on* I *and* $f(0) \leq 0$.

(ii) $f(K^*AK) \leq K^*f(A)K$ *for every contraction* K *and every Hermitian operator* A *with spectrum in* I.

(iii) $f(K_1^*AK_1 + K_2^*BK_2) \leq K_1^*f(A)K_1 + K_2^*f(B)K_2$ *for all operators* K_1, K_2 *such that* $K_1^*K_1 + K_2^*K_2 \leq I$ *and for all Hermitian* A, B *with spectrum in* I.

(iv) $f(PAP) \leq Pf(A)P$ *for all projections* P *and Hermitian operators* A *with spectrum in* I.

Proof. (i) $\Rightarrow$ (ii): Let $T = \left(\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and let U, V be the unitary operators defined in (V.6). Then

$$U^*TU = \begin{pmatrix} K^*AK & K^*AL \\ LAK & LAL \end{pmatrix}, \quad V^*TV = \begin{pmatrix} K^*AK & -K^*AL \\ -LAK & LAL \end{pmatrix}.$$

So,

$$\begin{pmatrix} K^*AK & 0 \\ 0 & LAL \end{pmatrix} = \frac{U^*TU + V^*TV}{2}.$$

Hence,

$$\begin{aligned} &\begin{pmatrix} f(K^*AK) & 0 \\ 0 & f(LAL) \end{pmatrix} \\ &= f\left(\frac{U^*TU + V^*TV}{2}\right) \\ &\leq \frac{f(U^*TU) + f(V^*TV)}{2} \\ &= \frac{U^*f(T)U + V^*f(T)V}{2} \\ &= \frac{1}{2}\left\{U^*\begin{pmatrix} f(A) & 0 \\ 0 & f(0) \end{pmatrix}U + V^*\begin{pmatrix} f(A) & 0 \\ 0 & f(0) \end{pmatrix}V\right\} \\ &\leq \frac{1}{2}\left\{U^*\begin{pmatrix} f(A) & 0 \\ 0 & 0 \end{pmatrix}U + V^*\begin{pmatrix} f(A) & 0 \\ 0 & 0 \end{pmatrix}V\right\} \\ &= \begin{pmatrix} K^*f(A)K & 0 \\ 0 & Lf(A)L \end{pmatrix}. \end{aligned}$$

Hence, $f(K^*AK) \leq K^*f(A)K$.

(ii) $\Rightarrow$ (iii): Let $T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$, $K = \begin{pmatrix} K_1 & 0 \\ K_2 & 0 \end{pmatrix}$. Then K is a contraction. Note that

$$K^*TK = \begin{pmatrix} K_1^*AK_1 + K_2^*BK_2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence,

$$\begin{pmatrix} f(K_1^*AK_1+K_2^*BK_2) & 0 \\ 0 & f(0) \end{pmatrix} = f(K^*TK) \leq K^*f(T)K$$
$$= \begin{pmatrix} K_1^*f(A)K_1+K_2^*f(B)K_2 & 0 \\ 0 & 0 \end{pmatrix}.$$

(iii) $\Rightarrow$ (iv) obviously.

(iv) $\Rightarrow$(i): Let A, B be Hermitian operators with spectrum in I and let $0 \leq \lambda \leq 1$. Let $T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$, $P = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ and let W be the unitary operator defined by (V.7). Then

$$PW^*TWP = \begin{pmatrix} \lambda A+(1-\lambda)B & 0 \\ 0 & 0 \end{pmatrix}.$$

So,

$$\begin{pmatrix} f(\lambda A+(1-\lambda)B) & 0 \\ 0 & f(0) \end{pmatrix} = f(PW^*TWP)$$
$$\leq Pf(W^*TW)P = PW^*f(T)WP$$
$$= \begin{pmatrix} \lambda f(A)+(1-\lambda)f(B) & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence, f is operator convex and $f(0) \leq 0$. ■

Exercise V.2.4 *(i) Let* λ_1, λ_2 *be positive real numbers such that* $\lambda_1\lambda_2 \geq C^*C$. *Then* $\begin{pmatrix} \lambda_1 I & C^* \\ C & \lambda_2 I \end{pmatrix}$ *is positive. (Use Proposition I.3.5.)*

(ii) Let $\begin{pmatrix} A & C^* \\ C & B \end{pmatrix}$ *be a Hermitian operator. Then for every* $\varepsilon > 0$, *there exists* $\lambda > 0$ *such that*

$$\begin{pmatrix} A & C^* \\ C & B \end{pmatrix} \leq \begin{pmatrix} A+\varepsilon I & 0 \\ 0 & \lambda I \end{pmatrix}.$$

The next two theorems are among the several results that describe the connections between operator convexity and operator monotonicity.

Theorem V.2.5 *Let f be a (continuous) function mapping the positive half-line* $[0,\infty)$ *into itself. Then f is operator monotone if and only if it is operator concave.*

Proof. Suppose f is operator monotone. If we show that $f(K^*AK) \geq K^*f(A)K$ for every positive operator A and contraction K, then it would follow from Theorem V.2.3 that f is operator concave. Let $T = \left(\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and let U be the unitary operator defined in (V.6). Then $U^*TU = \left(\begin{smallmatrix} K^*AK & K^*AL \\ LAK & LAL \end{smallmatrix}\right)$.

By the assertion in Exercise V.2.4(ii), given any $\varepsilon > 0$, there exists $\lambda > 0$ such that

$$U^*TU \leq \begin{pmatrix} K^*AK + \varepsilon & 0 \\ 0 & \lambda I \end{pmatrix}.$$

Replacing T by $f(T)$, we get

$$\begin{aligned} \begin{pmatrix} K^*f(A)K & K^*f(A)L \\ Lf(A)K & Lf(A)L \end{pmatrix} &= U^*f(T)U = f(U^*TU) \\ &\leq \begin{pmatrix} f(K^*AK + \varepsilon) & 0 \\ 0 & f(\lambda)I \end{pmatrix} \end{aligned}$$

by the operator monotonicity of f. In particular, this shows $K^*f(A)K \leq f(K^*AK + \varepsilon)$ for every $\varepsilon > 0$. Hence $K^*f(A)K \leq f(K^*AK)$.

Conversely, suppose f is operator concave. Let $0 \leq A \leq B$. Then for any $0 < \lambda < 1$ we can write

$$\lambda B = \lambda A + (1-\lambda)\frac{\lambda}{1-\lambda}(B - A).$$

Since f is operator concave, this gives

$$f(\lambda B) \geq \lambda f(A) + (1-\lambda) f\left(\frac{\lambda}{1-\lambda}(B-A)\right).$$

Since $f(X)$ is positive for every positive X, it follows that $f(\lambda B) \geq \lambda f(A)$. Now let $\lambda \to 1$. This shows $f(B) \geq f(A)$. So f is operator monotone. ■

Corollary V.2.6 *Let f be a continuous function from $(0, \infty)$ into itself. If f is operator monotone then the function $g(t) = \frac{1}{f(t)}$ is operator convex.*

Proof. Let A, B be positive operators. Since f is operator concave, $f\left(\frac{A+B}{2}\right) \geq \frac{f(A)+f(B)}{2}$. Since the map $X \to X^{-1}$ is order-reversing and convex on positive operators (see Proposition V.1.6 and Exercise V.1.15), this gives

$$\left[f\left(\frac{A+B}{2}\right)\right]^{-1} \leq \left[\frac{f(A)+f(B)}{2}\right]^{-1} \leq \frac{f(A)^{-1}+f(B)^{-1}}{2}.$$

This is the same as saying g is operator convex. ■

Exercise V.2.7 *Let I be an interval containing 0, and let f be a real function on I with $f(0) \leq 0$. Show that for every Hermitian operator A with spectrum in I and for every projection P*

$$f(PAP) \leq Pf(PAP) = Pf(PAP)P.$$

Exercise V.2.8 *Let f be a continuous real function on $[0,\infty)$. Then for all positive operators A and projections P*

$$f(A^{1/2}PA^{1/2})A^{1/2}P = A^{1/2}Pf(PAP).$$

(Prove this first, by induction, for $f(t) = t^n$. Then use the Weierstrass approximation theorem to show that this is true for all f.)

Theorem V.2.9 *Let f be a (continuous) real function on the interval $[0,\alpha)$. Then the following two conditions are equivalent:*
(i) f is operator convex and $f(0) \leq 0$.
(ii) The function $g(t) = f(t)/t$ is operator monotone on $(0,\alpha)$.

Proof. (i) $\Rightarrow$ (ii): Let $0 < A \leq B$. Then $0 < A^{1/2} \leq B^{1/2}$. Hence, $B^{-1/2}A^{1/2}$ is a contraction by Lemma V.1.7. Therefore, using Theorem V.2.3 we see that

$$f(A) = f(A^{1/2}B^{-1/2}BB^{-1/2}A^{1/2}) \leq A^{1/2}B^{-1/2}f(B)B^{-1/2}A^{1/2}.$$

From this, one obtains, using Lemma V.1.5,

$$A^{-1/2}f(A)A^{-1/2} \leq B^{-1/2}f(B)B^{-1/2}.$$

Since all functions of an operator commute with each other, this shows that $A^{-1}f(A) \leq B^{-1}f(B)$. Thus, g is operator monotone.

(ii) $\Rightarrow$(i): If $f(t)/t$ is monotone on $(0,\alpha)$ we must have $f(0) \leq 0$. We will show that f satisfies the condition (iv) of Theorem V.2.3. Let P be any projection and let A be any positive operator with spectrum in $(0,\alpha)$. Then there exists an $\varepsilon > 0$ such that $(1+\varepsilon)A$ has its spectrum in $(0,\alpha)$. Since $P+\varepsilon I \leq (1+\varepsilon)I$, we have $A^{1/2}(P+\varepsilon I)A^{1/2} \leq (1+\varepsilon)A$. So, by the operator monotonicity of g, we have

$$A^{-1/2}(P+\varepsilon I)^{-1}A^{-1/2}f(A^{1/2}(P+\varepsilon I)A^{1/2}) \leq (1+\varepsilon)^{-1}A^{-1}f((1+\varepsilon)A).$$

Multiply both sides on the right by $A^{1/2}(P+\varepsilon I)$ and on the left by its conjugate $(P+\varepsilon I)A^{1/2}$. This gives

$$A^{-1/2}f(A^{1/2}(P+\varepsilon I)A^{1/2})A^{1/2}(P+\varepsilon I) \leq (1+\varepsilon)^{-1}(P+\varepsilon I)f((1+\varepsilon)A)(P+\varepsilon I).$$

Let $\varepsilon \to 0$. This gives

$$A^{-1/2}f(A^{1/2}PA^{1/2})A^{1/2}P \leq Pf(A)P.$$

Use the identity in Exercise V.2.8 to reduce this to $Pf(PAP) \leq Pf(A)P$, and then use the inequality in Exercise V.2.7 to conclude that $f(PAP) \leq Pf(A)P$, as desired. ∎

As corollaries to the above results, we deduce the following statements about the power functions .

Theorem V.2.10 *On the positive half-line* $(0, \infty)$ *the functions* $f(t) = t^r$, *where* r *is a real number, are operator monotone if and only if* $0 \leq r \leq 1$.

Proof. If $0 \leq r \leq 1$, we know that $f(t) = t^r$ is operator monotone by Theorem V.1.9. If r is not in $[0,1]$, then the function $f(t) = t^r$ is not concave on $(0, \infty)$. Therefore, it cannot be operator monotone by Theorem V.2.5. ∎

Exercise V.2.11 *Consider the functions* $f(t) = t^r$ *on* $(0, \infty)$. *Use Theorems V.2.9 and V.2.10 to show that if* $r \geq 0$, *then* $f(t)$ *is operator convex if and only if* $1 \leq r \leq 2$. *Use Corollary V.2.6 to show that* $f(t)$, *is operator convex if* $-1 \leq r \leq 0$. *(We will see later that* $f(t)$ *is not operator convex for any other value of* r.*)*

Exercise V.2.12 *A function* f *from* $(0, \infty)$ *into itself is both operator monotone and operator convex if and only if it is of the form* $f(t) = \alpha + \beta t, \ \ \alpha, \beta \geq 0$.

Exercise V.2.13 *Show that the function* $f(t) = -t \log t$ *is operator concave on* $(0, \infty)$.

V.3 Smoothness Properties

Let I be the open interval $(-1, 1)$. Let f be a continuously differentiable function on I. Then we denote by $f^{[1]}$ the function on $I \times I$ defined as

$$\begin{aligned} f^{[1]}(\lambda, \mu) &= \frac{f(\lambda) - f(\mu)}{\lambda - \mu}, \quad \text{if } \lambda \neq \mu \\ f^{[1]}(\lambda, \lambda) &= f'(\lambda). \end{aligned}$$

The expression $f^{[1]}(\lambda, \mu)$ is called the **first divided difference** of f at (λ, μ).

If Λ is a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$, all of which are in I, we denote by $f^{[1]}(\Lambda)$ the $n \times n$ symmetric matrix whose (i, j)-entry is $f^{[1]}(\lambda_i, \lambda_j)$. If A is Hermitian and $A = U \Lambda U^*$, let $f^{[1]}(A) = U f^{[1]}(\Lambda) U^*$.

Now consider the induced map f on the set of Hermitian matrices with eigenvalues in I. Such matrices form an open set in the real vector space of all Hermitian matrices. The map f is called (Fréchet) **differentiable** at A if there exists a linear transformation $Df(A)$ on the space of Hermitian matrices such that for all H

$$\|f(A + H) - f(A) - Df(A)(H)\| = o(\|H\|). \tag{V.8}$$

The linear operator $Df(A)$ is then called the **derivative** of f at A. Basic rules of the Fréchet differential calculus are summarised in Chapter 10. If

f is differentiable at A, then

$$Df(A)(H) = \frac{d}{dt}\bigg|_{t=0} f(A+tH). \tag{V.9}$$

There is an interesting relationship between the derivative $Df(A)$ and the matrix $f^{[1]}(A)$. This is explored in the next few paragraphs.

Lemma V.3.1 *Let f be a polynomial function. Then for every diagonal matrix Λ and for every Hermitian matrix H,*

$$Df(\Lambda)(H) = f^{[1]}(\Lambda) \circ H, \tag{V.10}$$

where $\circ$ stands for the Schur-product of two matrices.

Proof. Both sides of (V.10) are linear in f. Therefore, it suffices to prove this for the powers $f(t) = t^p, p = 1, 2, 3, \ldots$ For such f, using (V.9) one gets

$$Df(\Lambda)(H) = \sum_{k=1}^{p} \Lambda^{k-1} H \Lambda^{p-k}.$$

This is a matrix whose (i,j)-entry is $\sum_{k=1}^{p} \lambda_i^{k-1} \lambda_j^{p-k} h_{ij}$. On the other hand, the (i,j)-entry of $f^{[1]}(\Lambda)$ is $\sum_{k=1}^{p} \lambda_i^{k-1} \lambda_j^{p-k}$. ■

Corollary V.3.2 *If $A = U\Lambda U^*$ and f is a polynomial function, then*

$$Df(A)(H) = U[f^{[1]}(\Lambda) \circ (U^*HU)]U^*. \tag{V.11}$$

Proof. Note that

$$\frac{d}{dt}\bigg|_{t=0} f(U\Lambda U^* + tH) = U\left[\frac{d}{dt}\bigg|_{t=0} f(\Lambda + tU^*HU)\right]U^*,$$

and use (V.10). ■

Theorem V.3.3 *Let $f \in C^1(I)$ and let A be a Hermitian matrix with all its eigenvalues in I. Then*

$$Df(A)(H) = f^{[1]}(A) \circ H, \tag{V.12}$$

where $\circ$ denotes the Schur-product in a basis in which A is diagonal.

Proof. Let $A = U\Lambda U^*$, where Λ is diagonal. We want to prove that

$$Df(A)(H) = U[f^{[1]}(\Lambda) \circ (U^*HU)]U^*. \tag{V.13}$$

This has been proved for all polynomials f. We will extend its validity to all $f \in C^1$ by a continuity argument.

Denote the right-hand side of (V.13) by $\mathcal{D}f(A)(H)$. For each f in C^1, $\mathcal{D}f(A)$ is a linear map on Hermitian matrices. We have

$$\|\mathcal{D}f(A)(H)\|_2 = \|f^{[1]}(\Lambda) \circ (U^*HU)\|_2.$$

All entries of the matrix $f^{[1]}(\Lambda)$ are bounded by $\max\limits_{|t|\leq\|A\|}|f'(t)|$. (Use the mean value theorem.) Hence

$$\|\mathcal{D}f(A)(H)\|_2 \leq \max_{|t|\leq\|A\|}|f'(t)| \; \|H\|_2. \tag{V.14}$$

Let H be a Hermitian matrix with norm so small that the eigenvalues of $A+H$ are in I. Let $[a,b]$ be a closed interval in I containing the eigenvalues of both A and $A+H$. Choose a sequence of polynomials f_n such that $f_n \to f$ and $f_n' \to f'$ uniformly on $[a,b]$. Let $\mathcal{L}$ be the line segment joining A and $A+H$ in the space of Hermitian matrices. Then, by the mean value theorem (for Fréchet derivatives), we have

$$\begin{aligned} &\|f_m(A+H) - f_n(A+H) - (f_m(A) - f_n(A))\| \\ \leq\; & \|H\| \sup_{X\in\mathcal{L}} \|Df_m(X) - Df_n(X)\| \\ =\; & \|H\| \sup_{X\in\mathcal{L}} \|\mathcal{D}f_m(X) - \mathcal{D}f_n(X)\|. \end{aligned} \tag{V.15}$$

This is so because we have already shown that $Df_n = \mathcal{D}f_n$ for the polynomial functions f_n.

Let ε be any positive real number. The inequality (V.14) ensures that there exists a positive integer n_0 such that for $m, n \geq n_0$ we have

$$\sup_{X\in\mathcal{L}} \|\mathcal{D}f_m(X) - \mathcal{D}f_n(X)\| \leq \frac{\varepsilon}{3} \tag{V.16}$$

and

$$\|\mathcal{D}f_n(A) - \mathcal{D}f(A)\| \leq \frac{\varepsilon}{3}. \tag{V.17}$$

Let $m \to \infty$ and use (V.15) and (V.16) to conclude that

$$\|f(A+H) - f(A) - (f_n(A+H) - f_n(A))\| \leq \frac{\varepsilon}{3}\|H\|. \tag{V.18}$$

If $\|H\|$ is sufficiently small, then by the definition of the Fréchet derivative, we have

$$\|f_n(A+H) - f_n(A) - \mathcal{D}f_n(A)(H)\| \leq \frac{\varepsilon}{3}\|H\|. \tag{V.19}$$

Now we can write, using the triangle inequality,

$$\begin{aligned} &\|f(A+H) - f(A) - \mathcal{D}f(A)(H)\| \\ \leq\; & \|f(A+H) - f(A) - (f_n(A+H) - f_n(A))\| \\ & + \|f_n(A+H) - f_n(A) - \mathcal{D}f_n(A)(H)\| \\ & + \|(\mathcal{D}f(A) - \mathcal{D}f_n(A))(H)\|, \end{aligned}$$

and then use (V.17), (V.18), and (V.19) to conclude that, for $\|H\|$ sufficiently small, we have

$$\|f(A+H) - f(A) - \mathcal{D}f(A)(H)\| \leq \varepsilon \|H\|.$$

But this says that $Df(A) = \mathcal{D}f(A)$. ■

Let $t \to A(t)$ be a C^1 map from the interval $[0,1]$ into the space of Hermitian matrices that have all their eigenvalues in I. Let $f \in C^1(I)$, and let $F(t) = f(A(t))$. Then, by the chain rule, $Df(t) = DF(A(t))(A'(t))$. Therefore, by the theorem above, we have

$$F(1) - F(0) = \int_0^1 f^{[1]}(A(t)) \circ A'(t)dt, \tag{V.20}$$

where for each t the Schur-product is taken in a basis that diagonalises $A(t)$.

Theorem V.3.4 *Let $f \in C^1(I)$. Then f is operator monotone on I if and only if, for every Hermitian matrix A whose eigenvalues are in I, the matrix $f^{[1]}(A)$ is positive.*

Proof. Let f be operator monotone, and let A be a Hermitian matrix whose eigenvalues are in I. Let H be the matrix all whose entries are 1. Then H is positive. So, $A + tH \geq A$ if $t \geq 0$. Hence, $f(A+tH) - f(A)$ is positive for small positive t. This implies that $Df(A)(H) \geq 0$. So, by Theorem V.3.3, $f^{[1]}(A) \circ H \geq 0$. But, for this special choice of H, this just says that $f^{[1]}(A) \geq 0$.

To prove the converse, let A, B be Hermitian matrices whose eigenvalues are in I, and let $B \geq A$. Let $A(t) = (1-t)A + tB$, $0 \leq t \leq 1$. Then $A(t)$ also has all its eigenvalues in I. So, by the hypothesis, $f^{[1]}(A(t)) \geq 0$ for all t. Note that $A'(t) = B - A \geq 0$, for all t. Since the Schur-product of two positive matrices is positive, $f^{[1]}(A(t)) \circ A'(t)$ is positive for all t. So, by (V.20), $f(B) - f(A) \geq 0$. ■

Lemma V.3.5 *If f is continuous and operator monotone on $(-1,1)$, then for each $-1 \leq \lambda \leq 1$ the function $g_\lambda(t) = (t+\lambda)f(t)$ is operator convex.*

Proof. We will prove this using Theorem V.2.9. First assume that f is continuous and operator monotone on $[-1,1]$. Then the function $f(t-1)$ is operator monotone on $[0,2)$. Let $g(t) = tf(t-1)$. Then $g(0) = 0$ and the function $g(t)/t$ is operator monotone on $(0,2)$. Hence, by Theorem V.2.9, $g(t)$ is operator convex on $[0,2)$. This implies that the function $h_1(t) = g(t+1) = (t+1)f(t)$ is operator convex on $[-1,1)$. Instead of $f(t)$, if the same argument is applied to the function $-f(-t)$, which is also operator

monotone on $[-1, 1]$, we see that the function $h_2(t) = -(t + 1)f(-t)$ is operator convex on $[-1, 1)$. Changing t to $-t$ preserves convexity. So the function $h_3(t) = h_2(-t) = (t - 1)f(t)$ is also operator convex. But for $|\lambda| \leq 1$, $g_\lambda(t) = \frac{1+\lambda}{2}h_1(t) + \frac{1-\lambda}{2}h_3(t)$ is a convex combination of h_1 and h_3. So g_λ is also operator convex.

Now, given f continuous and operator monotone on $(-1, 1)$, the function $f((1 - \varepsilon)t)$ is continuous and operator monotone on $[-1, 1]$ for each $\varepsilon > 0$. Hence, by the special case considered above, the function $(t + \lambda)f((1 - \varepsilon)t)$ is operator convex. Let $\varepsilon \to 0$, and conclude that the function $(t + \lambda)f(t)$ is operator convex. ■

The next theorem says that every operator monotone function on I is in the class C^1. Later on, we will see that it is actually in the class C^∞. (This is so even if we do not assume that it is continuous to begin with.) In the proof we make use of some differentiability properties of convex functions and smoothing techniques. For the reader's convenience, these are summarised in Appendices 1 and 2 at the end of the chapter.

Theorem V.3.6 *Every operator monotone function f on I is continuously differentiable.*

Proof. Let $0 < \varepsilon < 1$, and let f_ε be a regularisation of f of order ε. (See Appendix 2.) Then f_ε is a C^∞ function on $(-1+\varepsilon, 1-\varepsilon)$. It is also operator monotone. Let $\tilde{f}(t) = \lim_{\varepsilon \to 0} f_\varepsilon(t)$. Then $\tilde{f}(t) = \frac{1}{2}[f(t+) + f(t-)]$.

Let $g_\varepsilon(t) = (t+1)f_\varepsilon(t)$. Then, by Lemma V.3.5, g_ε is operator convex. Let $\tilde{g}(t) = \lim_{\varepsilon \to 0} g_\varepsilon(t)$. Then $\tilde{g}(t)$ is operator convex. But every convex function (on an open interval) is continuous. So $\tilde{g}(t)$ is continuous. Since $\tilde{g}(t) = (t + 1)\tilde{f}(t)$ and $t + 1 > 0$ on I, this means that $\tilde{f}(t)$ is continuous. Hence $\tilde{f}(t) = f(t)$. We thus have shown that f is continuous.

Let $g(t) = (t + 1)f(t)$. Then g is a convex function on I. So g is left and right differentiable and the one-sided derivatives satisfy the properties

$$g'_-(t) \leq g'_+(t), \quad \lim_{s \downarrow t} g'_\pm(s) = g'_+(t), \quad \lim_{s \uparrow t} g'_\pm(s) = g'_-(t). \tag{V.21}$$

But $g'_\pm(t) = f(t) + (t + 1)f'_\pm(t)$. Since $t + 1 > 0$, the derivatives $f'_\pm(t)$ also satisfy relations like (V.21).

Now let $A = \begin{pmatrix} s & 0 \\ 0 & t \end{pmatrix}$, $s, t \in (-1, 1)$. If ε is sufficiently small, s, t are in $(-1+\varepsilon, 1-\varepsilon)$. Since f_ε is operator monotone on this interval, by Theorem V.3.4, the matrix $f_\varepsilon^{[1]}(A)$ is positive. This implies that

$$\left(\frac{f_\varepsilon(s) - f_\varepsilon(t)}{s - t}\right)^2 \leq f'_\varepsilon(s)f'_\varepsilon(t).$$

Let $\varepsilon \to 0$. Since $f_\varepsilon \to f$ uniformly on compact sets, $f_\varepsilon(s) - f_\varepsilon(t)$ converges to $f(s) - f(t)$. Also, $f'_\varepsilon(t)$ converges to $\frac{1}{2}[f'_+(t) + f'_-(t)]$. Therefore, the

above inequality gives, in the limit, the inequality

$$\left(\frac{f(s)-f(t)}{s-t}\right)^2 \leq \frac{1}{4}[f'_+(s)+f'_-(s)][f'_+(t)+f'_-(t)].$$

Now let $s \downarrow t$, and use the fact that the derivatives of f satisfy relations like (V.21). This gives

$$[f'_+(t)]^2 \leq \frac{1}{4}[f'_+(t)+f'_+(t)][f'_+(t)+f'_-(t)],$$

which implies that $f'_+(t) = f'_-(t)$. Hence f is differentiable. The relations (V.21), which are satisfied by f too, show that f' is continuous. ■

Just as monotonicity of functions can be studied via first divided differences, convexity requires **second divided differences**. These are defined as follows. Let f be twice continuously differentiable on the interval I. Then $f^{[2]}$ is a function defined on $I \times I \times I$ as follows. If $\lambda_1, \lambda_2, \lambda_3$ are distinct

$$f^{[2]}(\lambda_1,\lambda_2,\lambda_3) = \frac{f^{[1]}(\lambda_1,\lambda_2) - f^{[1]}(\lambda_1,\lambda_3)}{\lambda_2-\lambda_3}.$$

For other values of $\lambda_1, \lambda_2, \lambda_3$, $f^{[2]}$ is defined by continuity; e.g.,

$$f^{[2]}(\lambda,\lambda,\lambda) = \frac{1}{2}f''(\lambda).$$

Exercise V.3.7 *Show that if $\lambda_1, \lambda_2, \lambda_3$ are distinct, then $f^{[2]}(\lambda_1,\lambda_2,\lambda_3)$ is the quotient of the two determinants*

$$\begin{vmatrix} f(\lambda_1) & f(\lambda_2) & f(\lambda_3) \\ \lambda_1 & \lambda_2 & \lambda_3 \\ 1 & 1 & 1 \end{vmatrix} \quad \textit{and} \quad \begin{vmatrix} \lambda_1^2 & \lambda_2^2 & \lambda_3^2 \\ \lambda_1 & \lambda_2 & \lambda_3 \\ 1 & 1 & 1 \end{vmatrix}.$$

Hence the function $f^{[2]}$ is symmetric in its three arguments.

Exercise V.3.8 *If $f(t) = t^m$, $m = 2, 3, \ldots$, show that*

$$f^{[2]}(\lambda_1,\lambda_2,\lambda_3) = \sum_{\substack{0 \leq p,q,r \\ p+q+r=m-2}} \lambda_1^p \lambda_2^q \lambda_3^r.$$

Exercise V.3.9 *(i) Let $f(t) = t^m, m \geq 2$. Let A be an $n \times n$ diagonal matrix; $A = \sum_{i=1}^{n} \lambda_i P_i$, where P_i are the projections onto the coordinate axes. Show that for every H*

$$\begin{aligned} \left.\frac{d^2}{dt^2}\right|_{t=0} f(A+tH) &= 2 \sum_{p+q+r=m-2} A^p H A^q H A^r \\ &= 2 \sum_{p+q+r=m-2} \; \sum_{1 \leq i,j,k \leq n} \lambda_i^p \lambda_j^q \lambda_k^r P_i H P_j H P_k, \end{aligned}$$

and

$$\left.\frac{d^2}{dt^2}\right|_{t=0} f(A+tH) = 2\sum_{i,j,k} f^{[2]}(\lambda_i,\lambda_j,\lambda_k)P_iHP_jHP_k. \tag{V.22}$$

(ii) Use a continuity argument, like the one used in the proof of Theorem V.3.3, to show that this last formula is valid for all C^2 functions f.

Theorem V.3.10 *If $f \in C^2(I)$ and f is operator convex, then for each $\mu \in I$ the function $g(\lambda) = f^{[1]}(\mu,\lambda)$ is operator monotone.*

Proof. Since f is in the class C^2, g is in the class C^1. So, by Theorem V.3.4, it suffices to prove that, for each n, the $n \times n$ matrix with entries $g^{[1]}(\lambda_i,\lambda_j)$ is positive for all $\lambda_1,\ldots,\lambda_n$ in I.

Fix n and choose any $\lambda_1,\ldots,\lambda_{n+1}$ in I. Let A be the diagonal matrix with entries $\lambda_1,\ldots\lambda_{n+1}$. Since f is operator convex and is twice differentiable, for every Hermitian matrix H, the matrix $\left.\frac{d^2}{dt^2}\right|_{t=0} f(A+tH)$ must be positive. If we write $P_1,\ldots,P_{n+1}$ for the projections onto the coordinate axes, we have an explicit expression for this second derivative in (V.22). Choose H to be of the form

$$H = \begin{pmatrix} 0 & 0 & \cdots & & \bar{\xi}_1 \\ 0 & 0 & \cdots & & \bar{\xi}_2 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \xi_1 & \xi_2 & \cdots & \xi_n & 0 \end{pmatrix},$$

where $\xi_1,\ldots,\xi_n$ are any complex numbers. Let x be the $(n+1)$-vector $(1,1,\ldots,1,0)$. Then

$$\langle x, P_iHP_jHP_kx\rangle = \xi_k\bar{\xi}_i\delta_{j,n+1} \tag{V.23}$$

for $1 \le i,j,k \le n+1$, where $\delta_{j,n+1}$ is equal to 1 if $j = n+1$, and is equal to 0 otherwise. So, using the positivity of the matrix (V.22) and then (V.23), we have

$$\begin{aligned} 0 &\le \sum_{1\le i,j,k\le n+1} f^{[2]}(\lambda_i,\lambda_j,\lambda_k)\langle x, P_iHP_jHP_kx\rangle \\ &= \sum_{1\le i,k\le n} f^{[2]}(\lambda_i,\lambda_{n+1},\lambda_k)\xi_k\bar{\xi}_i. \end{aligned}$$

But,

$$\begin{aligned} f^{[2]}(\lambda_i,\lambda_{n+1},\lambda_k) &= \frac{f^{[1]}(\lambda_{n+1},\lambda_i) - f^{[1]}(\lambda_{n+1},\lambda_k)}{\lambda_i - \lambda_k} \\ &= g^{[1]}(\lambda_i,\lambda_k) \end{aligned}$$

(putting $\lambda_{n+1} = \mu$ in the definition of g). So we have

$$0 \leq \sum_{1 \leq i,k \leq n} g^{[1]}(\lambda_i, \lambda_k)\xi_k \, \bar{\xi}_i.$$

Since ξ_i are arbitrary complex numbers, this is equivalent to saying that the $n \times n$ matrix $[g^{[1]}(\lambda_i, \lambda_k)]$ is positive. ■

Corollary V.3.11 *If $f \in C^2(I)$, $f(0) = 0$, and f is operator convex, then the function $g(t) = \frac{f(t)}{t}$ is operator monotone.*

Proof. By the theorem above, the function $f^{[1]}(0, t)$ is operator monotone. But this is just the function $f(t)/t$ in this case. ■

Corollary V.3.12 *If f is operator monotone on I and $f(0) = 0$, then the function $g(t) = \frac{t+\lambda}{t} f(t)$ is operator monotone for $|\lambda| \leq 1$.*

Proof. First assume that $f \in C^2(I)$. By Lemma V.3.5, the function $g_\lambda(t) = (t + \lambda)f(t)$ is operator convex. By Corollary V.3.11, therefore, $g(t)$ is operator monotone.

If f is not in the class C^2, consider its regularisations f_ε. These are in C^2. Apply the special case of the above paragraph to the functions $f_\varepsilon(t) - f_\varepsilon(0)$, and then let $\varepsilon \to 0$. ■

Corollary V.3.13 *If f is operator monotone on I and $f(0) = 0$, then f is twice differentiable at 0.*

Proof. By Corollary V.3.12, the function $g(t) = (1 + \frac{1}{t})f(t)$ is operator monotone, and by Theorem V.3.6, it is continuously differentiable. So the function h defined as $h(t) = \frac{1}{t}f(t)$, $h(0) = f'(0)$ is continuously differentiable. This implies that f is twice differentiable at 0. ■

Exercise V.3.14 *Let f be a continuous operator monotone function on I. Then the function $F(t) = \int_0^t f(s)ds$ is operator convex.*

Exercise V.3.15 *Let $f \in C^1(I)$. Then f is operator convex if and only if for all Hermitian matrices A, B with eigenvalues in I we have*

$$f(A) - f(B) \geq f^{[1]}(B) \circ (A - B),$$

where $\circ$ denotes the Schur-product in a basis in which B is diagonal.

V.4 Loewner's Theorems

Consider all functions f on the interval $I = (-1,1)$ that are operator monotone and satisfy the conditions

$$f(0) = 0, \qquad f'(0) = 1. \tag{V.24}$$

Let K be the collection of all such functions. Clearly, K is a convex set. We will show that this set is compact in the topology of pointwise convergence and will find its extreme points. This will enable us to write an integral representation for functions in K.

Lemma V.4.1 *If $f \in K$, then*

$$\begin{aligned} f(t) &\leq \frac{t}{1-t} \quad \textit{for} \quad 0 \leq t < 1, \\ f(t) &\geq \frac{t}{1+t} \quad \textit{for} -1 < t < 0, \\ |f''(0)| &\leq 2. \end{aligned}$$

Proof. Let $A = \begin{pmatrix} t & 0 \\ 0 & 0 \end{pmatrix}$. By Theorem V.3.4, the matrix

$$f^{[1]}(A) = \begin{pmatrix} f'(t) & f(t)/t \\ f(t)/t & 1 \end{pmatrix}$$

is positive. Hence,

$$\frac{f(t)^2}{t^2} \leq f'(t). \tag{V.25}$$

Let $g_{\pm}(t) = (t \pm 1)f(t)$. By Lemma V.3.5, both functions $g_{\pm}$ are convex. Hence their derivatives are monotonically increasing functions. Since $g'_{\pm}(t) = f(t) + (t \pm 1)f'(t)$ and $g'_{\pm}(0) = \pm 1$, this implies that

$$f(t) + (t-1)f'(t) \geq -1 \quad \text{for} \quad t > 0 \tag{V.26}$$

and

$$f(t) + (t+1)f'(t) \leq 1 \quad \text{for} \quad t < 0. \tag{V.27}$$

From (V.25) and (V.26) we obtain

$$f(t) + 1 \geq \frac{(1-t)f(t)^2}{t^2} \quad \text{for} \quad t > 0. \tag{V.28}$$

Now suppose that for some $0 < t < 1$ we have $f(t) > \frac{t}{1-t}$. Then $f(t)^2 > \frac{t}{1-t}f(t)$. So, from (V.28), we get $f(t)+1 > \frac{f(t)}{t}$. But this gives the inequality $f(t) < \frac{t}{1-t}$, which contradicts our assumption. This shows that $f(t) \leq \frac{t}{1-t}$ for $0 \leq t < 1$. The second inequality of the lemma is obtained by the same argument using (V.27) instead of (V.26).

We have seen in the proof of Corollary V.3.13 that

$$f'(0) + \frac{1}{2}f''(0) = \lim_{t\to 0} \frac{(1+t^{-1})f(t) - f'(0)}{t}.$$

Let $t \downarrow 0$ and use the first inequality of the lemma to conclude that this limit is smaller than 2. Let $t \uparrow 0$, and use the second inequality to conclude that it is bigger than 0. Together, these two imply that $|f''(0)| \le 2$. ■

Proposition V.4.2 *The set K is compact in the topology of pointwise convergence.*

Proof. Let $\{f_i\}$ be any net in K. By the lemma above, the set $\{f_i(t)\}$ is bounded for each t. So, by Tychonoff's Theorem, there exists a subnet $\{f_i\}$ that converges pointwise to a bounded function f. The limit function f is operator monotone, and $f(0) = 0$. If we show that $f'(0) = 1$, we would have shown that $f \in K$, and hence that K is compact.

By Corollary V.3.12, each of the functions $(1 + \frac{1}{t})f_i(t)$ is monotone on $(-1, 1)$. Since for all i, $\lim_{t\to 0}(1 + \dfrac{1}{t})f_i(t) = f_i'(0) = 1$, we see that $(1 + \frac{1}{t})f_i(t) \ge 1$ if $t \ge 0$ and is ≤ 1 if $t \le 0$. Hence, if $t > 0$, we have $(1 + \frac{1}{t})f(t) \ge 1$; and if $t < 0$, we have the opposite inequality. Since f is continuously differentiable, this shows that $f'(0) = 1$. ■

Proposition V.4.3 *All extreme points of the set K have the form*

$$f(t) = \frac{t}{1-\alpha t}, \quad \textit{where} \quad \alpha = \frac{1}{2}f''(0).$$

Proof. Let $f \in K$. For each λ, $-1 < \lambda < 1$, let

$$g_\lambda(t) = (1 + \frac{\lambda}{t})f(t) - \lambda.$$

By Corollary V.3.12, g_λ is operator monotone. Note that $g_\lambda(0) = 0$, since $f(0) = 0$ and $f'(0) = 1$. Also, $g_\lambda'(0) = 1 + \frac{1}{2}\lambda f''(0)$. So the function h_λ defined as

$$h_\lambda(t) = \frac{1}{1 + \frac{1}{2}\lambda f''(0)}[(1 + \frac{\lambda}{t})f(t) - \lambda]$$

is in K. Since $|f''(0)| \le 2$, we see that $|\frac{1}{2}\lambda f''(0)| < 1$. We can write

$$f = \frac{1}{2}(1 + \frac{1}{2}\lambda f''(0))h_\lambda + \frac{1}{2}(1 - \frac{1}{2}\lambda f''(0))h_{-\lambda}.$$

So, if f is an extreme point of K, we must have $f = h_\lambda$. This says that

$$(1 + \frac{1}{2}\lambda f''(0))f(t) = (1 + \frac{\lambda}{t})f(t) - \lambda,$$

from which we can conclude that

$$f(t) = \frac{t}{1 - \frac{1}{2}f''(0)t}.$$

■

Theorem V.4.4 *For each f in K there exists a unique probability measure μ on $[-1, 1]$ such that*

$$f(t) = \int_{-1}^{1} \frac{t}{1 - \lambda t} d\mu(\lambda). \tag{V.29}$$

Proof. For $-1 \le \lambda \le 1$, consider the functions $h_\lambda(t) = \frac{t}{1-\lambda t}$. By Proposition V.4.3, the extreme points of K are included in the family $\{h_\lambda\}$. Since K is compact and convex, it must be the closed convex hull of its extreme points. (This is the Krein-Milman Theorem.) Finite convex combinations of elements of the family $\{h_\lambda : -1 \le \lambda \le 1\}$ can also be written as $\int h_\lambda d\nu(\lambda)$, where ν is a probability measure on $[-1, 1]$ with finite support. Since f is in the closure of these combinations, there exists a net $\{\nu_i\}$ of finitely supported probability measures on $[-1, 1]$ such that the net $f_i(t) = \int h_\lambda(t) d\nu_i(\lambda)$ converges to $f(t)$. Since the space of the probability measures is weak* compact, the net ν_i has an accumulation point μ. In other words, a subnet of $\int h_\lambda d\nu_i(\lambda)$ converges to $\int h_\lambda d\mu(\lambda)$. So $f(t) = \int h_\lambda(t) d\mu(\lambda) = \int \frac{t}{1-\lambda t} d\mu(\lambda)$.

Now suppose that there are two measures μ_1 and μ_2 for which the representation (V.29) is valid. Expand the integrand as a power series $\frac{t}{1-\lambda t} = \sum_{n=0}^{\infty} t^{n+1} \lambda^n$ convergent uniformly in $|\lambda| \le 1$ for every fixed t with $|t| < 1$. This shows that

$$\sum_{n=0}^{\infty} t^{n+1} \int_{-1}^{1} \lambda^n d\mu_1(\lambda) = \sum_{n=0}^{\infty} t^{n+1} \int_{-1}^{1} \lambda^n d\mu_2(\lambda)$$

for all $|t| < 1$. The identity theorem for power series now shows that

$$\int_{-1}^{1} \lambda^n d\mu_1(\lambda) = \int_{-1}^{1} \lambda^n d\mu_2(\lambda), \quad n = 0, 1, 2, \ldots$$

But this is possible if and only if $\mu_1 = \mu_2$. ■

One consequence of the uniqueness of the measure μ in the representation (V.29) is that every function h_{λ_0} is an extreme point of K (because it can be represented as an integral like this with μ concentrated at λ_0).

The normalisations (V.24) were required to make the set K compact. They can now be removed. We have the following result.

Corollary V.4.5 *Let f be a nonconstant operator monotone function on $(-1,1)$. Then there exists a unique probability measure μ on $[-1,1]$ such that*

$$f(t) = f(0) + f'(0) \int_{-1}^{1} \frac{t}{1-\lambda t} d\mu(\lambda). \tag{V.30}$$

Proof. Since f is monotone and is not a constant, $f'(0) \neq 0$. Now note that the function $\frac{f(t)-f(0)}{f'(0)}$ is in K. ■

It is clear from the representation (V.30) that every operator monotone function on $(-1,1)$ is infinitely differentiable. Hence, by the results of earlier sections, every operator convex function is also infinitely differentiable.

Theorem V.4.6 *Let f be a nonlinear operator convex function on $(-1,1)$. Then there exists a unique probability measure μ on $[-1,1]$ such that*

$$f(t) = f(0) + f'(0)t + \frac{1}{2} f''(0) \int_{-1}^{1} \frac{t^2}{1-\lambda t} d\mu(\lambda). \tag{V.31}$$

Proof. Assume, without loss of generality, that $f(0) = 0$ and $f'(0) = 0$. Let $g(t) = f(t)/t$. Then g is operator monotone by Corollary V.3.11, $g(0) = 0$, and $g'(0) = \frac{1}{2} f''(0)$. So g has a representation like (V.30), from which the representation (V.31) for f follows. ■

We have noted that the integral representation (V.30) implies that every operator monotone function on $(-1,1)$ is infinitely differentiable. In fact, we can conclude more. This representation shows that f has an analytic continuation

$$f(z) = f(0) + f'(0) \int_{-1}^{1} \frac{z}{1-\lambda z} d\mu(\lambda) \tag{V.32}$$

defined everywhere on the complex plane except on $(-\infty,-1] \cup [1,\infty)$. Note that

$$\text{Im} \frac{z}{1-\lambda z} = \frac{\text{Im } z}{|1-\lambda z|^2}.$$

So f defined above maps the upper half-plane $H_+ = \{z : \text{Im } z > 0\}$ into itself. It also maps the lower half-plane H_- into itself. Further, $f(\bar{z}) = \overline{f(z)}$. In other words, the function f on H_- is an analytic continuation of f on H_+ across the interval $(-1,1)$ obtained by reflection.

This is a very important observation, because there is a very rich theory of analytic functions in a half-plane that we can exploit now. Before doing so, let us now do away with the special interval $(-1,1)$. Note that a function f is operator monotone on an interval (a,b) if and only if the function

$f(\frac{(b-a)t}{2} + \frac{b+a}{2})$ is operator monotone on $(-1,1)$. So, all results obtained for operator monotone functions on $(-1,1)$ can be extended to functions on (a,b). We have proved the following.

Theorem V.4.7 *If f is an operator monotone function on (a,b), then f has an analytic continuation to the upper half-plane H_+ that maps H_+ into itself. It also has an analytic continuation to the lower-half plane H_-, obtained by reflection across (a,b).*

The converse of this is also true: if a real function f on (a,b) has an analytic continuation to H_+ mapping H_+ into itself, then f is operator monotone on (a,b). This is proved below.

Let P be the class of all complex analytic functions defined on H_+ with their ranges in the closed upper half-plane $\{z : \text{Im } z \geq 0\}$. This is called the class of **Pick functions**. Since every nonconstant analytic function is an open map, if f is a nonconstant Pick function, then the range of f is contained in H_+. It is obvious that P is a convex cone, and the composition of two nonconstant functions in P is again in P.

Exercise V.4.8 *(i) For $0 \leq r \leq 1$, the function $f(z) = z^r$ is in P.*

(ii) The function $f(z) = \log z$ is in P.

(iii) The function $f(z) = \tan z$ is in P.

(iv) The function $f(z) = -\frac{1}{z}$ is in P.

(v) If f is in P, then so is the function $\frac{-1}{f}$.

Given any open interval (a,b), let $P(a,b)$ be the class of Pick functions that admit an analytic continuation across (a,b) into the lower half-plane and the continuation is by reflection. In particular, such functions take only real values on (a,b), and if they are nonconstant, they assume real values only on (a,b). The set $P(a,b)$ is a convex cone.

Let $f \in P(a,b)$ and write $f(z) = u(z) + iv(z)$, where as usual $u(z)$ and $v(z)$ denote the real and imaginary parts of f. Since $v(x) = 0$ for $a < x < b$, we have $v(x+iy) - v(x) \geq 0$ if $y > 0$. This implies that the partial derivative $v_y(x) \geq 0$ and hence, by the Cauchy-Riemann equations, $u_x(x) \geq 0$. Thus, on the interval (a,b), $f(x) = u(x)$ is monotone. In fact, we will soon see that f is operator monotone on (a,b). This is a consequence of a theorem of Nevanlinna that gives an integral representation of Pick functions. We will give a proof of this now using some elementary results from Fourier analysis. The idea is to use the conformal equivalence between H_+ and the unit disk D to transfer the problem to D, and then study the real part u of f. This is a harmonic function on D, so we can use standard facts from Fourier analysis.

Theorem V.4.9 *Let u be a nonnegative harmonic function on the unit disk $D = \{z : |z| < 1\}$. Then there exists a finite measure m on $[0, 2\pi]$ such*

that

$$u(re^{i\theta}) = \int_0^{2\pi} \frac{1-r^2}{1+r^2-2r\,\cos(\theta-t)}dm(t). \tag{V.33}$$

Conversely, any function of this form is positive and harmonic on the unit disk D.

Proof. Let u be any continuous real function defined on the closed unit disk that is harmonic in D. Then, by a well-known and elementary theorem in analysis,

$$\begin{aligned} u(re^{i\theta}) &= \frac{1}{2\pi}\int_0^{2\pi} \frac{1-r^2}{1+r^2-2r\,\cos(\theta-t)}u(e^{it})dt \\ &= \frac{1}{2\pi}\int_0^{2\pi} P_r(\theta-t)u(e^{it})dt, \end{aligned} \tag{V.34}$$

where $P_r(\theta)$ is the Poisson kernel (defined by the above equation) for $0 \leq r < 1$, $0 \leq \theta \leq 2\pi$. If u is nonnegative, put $dm(t) = \frac{1}{2\pi}u(e^{it})dt$. Then m is a positive measure on $[0, 2\pi]$. By the mean value property of harmonic functions, the total mass of this measure is

$$\frac{1}{2\pi}\int_0^{2\pi} u(e^{it})dt = u(0). \tag{V.35}$$

So we do have a representation of the form (V.33) under the additional hypothesis that u is continuous on the closed unit disk.

The general case is a consequence of this. Let u be positive and harmonic in D. Then, for $\varepsilon > 0$, the function $u_\varepsilon(z) = u(\frac{z}{1+\varepsilon})$ is positive and harmonic in the disk $|z| < 1+\varepsilon$. Therefore, it can be represented in the form (V.33) with a measure $m_\varepsilon(t)$ of finite total mass $u_\varepsilon(0) = u(0)$. As $\varepsilon \to 0$, u_ε converges to u uniformly on compact subsets of D. Since the measures m_ε all have the same mass, using the weak* compactness of the space of probability measures, we conclude that there exists a positive measure m such that

$$u(re^{i\theta}) = \lim_{\varepsilon\to 0} u_\varepsilon(re^{i\theta}) = \int_0^{2\pi} \frac{1-r^2}{1+r^2-2r\,\cos(\theta-t)}dm(t).$$

Conversely, since the Poisson kernel P_r is nonnegative any function represented by (V.33) is nonnegative. ∎

Theorem V.4.9 is often called the **Herglotz Theorem**. It says that *every nonnegative harmonic function on the unit disk is the Poisson integral of a positive measure.*

Recall that two harmonic functions u, v are called **harmonic conjugates** if the function $f(z) = u(z) + iv(z)$ is analytic. Every harmonic function u has a harmonic conjugate that is uniquely determined up to an additive constant.

Theorem V.4.10 *Let $f(z) = u(z) + iv(z)$ be analytic on the unit disk D. If $u(z) \geq 0$, then there exists a finite positive measure m on $[0, 2\pi]$ such that*

$$f(z) = \int_0^{2\pi} \frac{e^{it} + z}{e^{it} - z}\, dm(t) + iv(0). \tag{V.36}$$

Conversely, every function of this form is analytic on D and has a positive real part.

Proof. By Theorem V.4.9, the function u can be written as in (V.33). The Poisson kernel P_r, $0 \leq r < 1$, can be written as

$$P_r(\theta) = \frac{1 - r^2}{1 + r^2 - 2r\cos\theta} = \sum_{-\infty}^{\infty} r^{|n|} e^{in\theta} = \operatorname{Re} \frac{1 + re^{i\theta}}{1 - re^{i\theta}}.$$

Hence,

$$P_r(\theta - t) = \operatorname{Re} \frac{1 + re^{i(\theta - t)}}{1 - re^{i(\theta - t)}} = \operatorname{Re} \frac{e^{it} + re^{i\theta}}{e^{it} - re^{i\theta}},$$

and

$$u(z) = \operatorname{Re} \int_0^{2\pi} \frac{e^{it} + z}{e^{it} - z}\, dm(t).$$

So, $f(z)$ differs from this last integral only by an imaginary constant. Putting $z = 0$, one sees that this constant is $iv(0)$.

The converse statement is easy to prove. ■

Next, note that the disk D and the half-plane H_+ are conformally equivalent,, i.e., there exists an analytic isomorphism between these two spaces. For $z \in D$, let

$$\zeta(z) = \frac{1}{i}\, \frac{z + 1}{z - 1}. \tag{V.37}$$

Then $\zeta \in H_+$. The inverse of this map is given by

$$z(\zeta) = \frac{\zeta - i}{\zeta + i}. \tag{V.38}$$

Using these transformations, we can establish an equivalence between the class P and the class of analytic functions on D with positive real part. If f is a function in the latter class, let

$$\varphi(\zeta) = if(z(\zeta)). \tag{V.39}$$

Then $\varphi \in P$. The inverse of this transformation is

$$f(z) = -i\varphi(\zeta(z)). \tag{V.40}$$

Using these ideas we can prove the following theorem, called **Nevanlinna's Theorem.**

Theorem V.4.11 *A function φ is in the Pick class if and only if it has a representation*

$$\varphi(\zeta) = \alpha + \beta\zeta + \int_{-\infty}^{\infty} \frac{1+\lambda\zeta}{\lambda - \zeta} d\nu(\lambda), \tag{V.41}$$

where α is a real number, $\beta \geq 0$, and ν is a positive finite measure on the real line.

Proof. Let f be the function on D associated with φ via the transformation (V.40). By Theorem V.4.10, there exists a finite positive measure m on $[0, 2\pi]$ such that

$$f(z) = \int_0^{2\pi} \frac{e^{it}+z}{e^{it}-z} dm(t) - i\alpha.$$

If $f(z) = u(z) + iv(z)$, then $\alpha = -v(0)$, and the total mass of m is $u(0)$. If the measure m has a positive mass at the singleton $\{0\}$, let this mass be β. Then the expression above reduces to

$$f(z) = \int_{(0,2\pi)} \frac{e^{it}+z}{e^{it}-z}\, dm(t) + \beta\, \frac{1+z}{1-z} - i\alpha.$$

Using the transformations (V.38) and (V.39), we get from this

$$\varphi(\zeta) = \alpha + \beta\zeta + i \int_{(0,2\pi)} \frac{e^{it} + \frac{\zeta - i}{\zeta + i}}{e^{it} - \frac{\zeta - i}{\zeta + i}} dm(t).$$

The last term above is equal to

$$\int_{(0,2\pi)} \frac{\zeta\, \cos\frac{t}{2} - \sin\frac{t}{2}}{\zeta\, \sin\frac{t}{2} + \cos\frac{t}{2}} dm(t).$$

Now, introduce a change of variables $\lambda = -\cot\frac{t}{2}$. This maps $(0, 2\pi)$ onto $(-\infty, \infty)$. The measure m is transformed by the above map to a finite measure ν on $(-\infty, \infty)$ and the above integral is transformed to

$$\int_{-\infty}^{\infty} \frac{1+\lambda\zeta}{\lambda - \zeta}\, d\nu(\lambda).$$

This shows that φ can be represented in the form (V.41).

It is easy to see that every function of this form is a Pick function. ■

There is another form in which it is convenient to represent Pick functions. Note that

$$\frac{1+\lambda\zeta}{\lambda-\zeta}=\left(\frac{1}{\lambda-\zeta}-\frac{\lambda}{\lambda^2+1}\right)(\lambda^2+1).$$

So, if we write $d\mu(\lambda) = (\lambda^2+1)d\nu(\lambda)$, then we obtain from (V.41) the representation

$$\varphi(\zeta)=\alpha+\beta\zeta+\int_{-\infty}^{\infty}\left[\frac{1}{\lambda-\zeta}-\frac{\lambda}{\lambda^2+1}\right]d\mu(\lambda), \tag{V.42}$$

where μ is a positive Borel measure on $\mathbb{R}$, for which $\int \frac{1}{\lambda^2+1}d\mu(\lambda)$ is finite. (A Borel measure on $\mathbb{R}$ is a measure defined on Borel sets that puts finite mass on bounded sets.)

Now we turn to the question of uniqueness of the above representations. It is easy to see from (V.41) that

$$\alpha = \text{Re}\ \varphi(i). \tag{V.43}$$

Therefore, α is uniquely determined by φ. Now let η be any positive real number. From (V.41) we see that

$$\frac{\varphi(i\eta)}{i\eta}=\frac{\alpha}{i\eta}+\beta+\int_{-\infty}^{\infty}\frac{1+\lambda^2+i\lambda(\eta-\eta^{-1})}{\lambda^2+\eta^2}d\nu(\lambda).$$

As $\eta\to\infty$, the integrand converges to 0 for each λ. The real and imaginary parts of the integrand are uniformly bounded by 1 when $\eta>1$. So by the Lebesgue Dominated Convergence Theorem, the integral converges to 0 as $\eta\to\infty$. Thus,

$$\beta=\lim_{\eta\to\infty}\varphi(i\eta)/i\eta, \tag{V.44}$$

and thus β is uniquely determined by φ.

Now we will prove that the measure $d\mu$ in (V.42), is uniquely determined by φ. Denote by μ the unique right continuous monotonically increasing function on $\mathbb{R}$ satisfying $\mu(0)=0$ and $\mu((a,b])=\mu(b)-\mu(a)$ for every interval $(a,b]$. (This is called the **distribution function** associated with $d\mu$.) We will prove the following result, called the **Stieltjes inversion formula**, from which it follows that μ is unique.

Theorem V.4.12 *If the Pick function φ is represented by (V.42), then for any a,b that are points of continuity of the distribution function μ we have*

$$\mu(b)-\mu(a)=\lim_{\eta\to 0}\frac{1}{\pi}\int_a^b \text{Im}\ \varphi(x+i\eta)dx. \tag{V.45}$$

Proof. From (V.42) we see that

$$\frac{1}{\pi}\int_a^b \operatorname{Im}\varphi(x+i\eta)dx = \frac{1}{\pi}\int_a^b \left[\beta\eta + \int_{-\infty}^{\infty} \frac{\eta}{(\lambda-x)^2+\eta^2}d\mu(\lambda)\right] dx$$
$$= \frac{1}{\pi}\left[\beta\eta(b-a) + \int_{-\infty}^{\infty}\int_a^b \frac{\eta dx}{(x-\lambda)^2+\eta^2}d\mu(\lambda)\right],$$

the interchange of integrals being permissible by Fubini's Theorem. As $\eta \to 0$, the first term in the square brackets above goes to 0. The inner integral can be calculated by the change of variables $u = \frac{x-\lambda}{\eta}$. This gives

$$\int_a^b \frac{\eta dx}{(x-\lambda)^2+\eta^2} = \int_{\frac{a-\lambda}{\eta}}^{\frac{b-\lambda}{\eta}} \frac{du}{u^2+1}$$
$$= \arctan\left(\frac{b-\lambda}{\eta}\right) - \arctan\left(\frac{a-\lambda}{\eta}\right).$$

So to prove (V.45), we have to show that

$$\mu(b)-\mu(a) = \lim_{\eta\to 0}\frac{1}{\pi}\int_{-\infty}^{\infty}\left[\arctan\left(\frac{b-\lambda}{\eta}\right) - \arctan\left(\frac{a-\lambda}{\eta}\right)\right] d\mu(\lambda).$$

We will use the following properties of the function arctan. This is a monotonically increasing odd function on $(-\infty,\infty)$ whose range is $(-\frac{\pi}{2}, \frac{\pi}{2})$. So,

$$0 \le \arctan\left(\frac{b-\lambda}{\eta}\right) - \arctan\left(\frac{a-\lambda}{\eta}\right) \le \pi.$$

If $(b-\lambda)$ and $(a-\lambda)$ have the same sign, then by the addition law for arctan we have,

$$\arctan\left(\frac{b-\lambda}{\eta}\right) - \arctan\left(\frac{a-\lambda}{\eta}\right) = \arctan\frac{\eta(b-a)}{\eta^2+(b-\lambda)(a-\lambda)}.$$

If x is positive, then

$$\arctan x = \int_0^x \frac{dt}{1+t^2} \le \int_0^x dt = x.$$

Now, let ε be any given positive number. Since a and b are points of continuity of μ, we can choose δ such that

$$\mu(a+\delta)-\mu(a-\delta) \le \varepsilon/5,$$
$$\mu(b+\delta)-\mu(b-\delta) \le \varepsilon/5.$$

We then have,

$$
\begin{aligned}
&\left| \mu(b)-\mu(a)-\frac{1}{\pi}\int_{-\infty}^{\infty}\left[\arctan\left(\frac{b-\lambda}{\eta}\right)-\arctan\left(\frac{a-\lambda}{\eta}\right)\right]d\mu(\lambda)\right| \\
\le\ & \frac{1}{\pi}\int_{b}^{\infty}\left[\arctan\left(\frac{b-\lambda}{\eta}\right)-\arctan\left(\frac{a-\lambda}{\eta}\right)\right]d\mu(\lambda) \\
&+\frac{1}{\pi}\int_{a}^{b}\left[\pi-\arctan\left(\frac{b-\lambda}{\eta}\right)+\arctan\left(\frac{a-\lambda}{\eta}\right)\right]d\mu(\lambda) \\
&+\frac{1}{\pi}\int_{-\infty}^{a}\left[\arctan\left(\frac{b-\lambda}{\eta}\right)-\arctan\left(\frac{a-\lambda}{\eta}\right)\right]d\mu(\lambda) \\
\le\ & \frac{2\varepsilon}{5}+\frac{1}{\pi}\int_{b+\delta}^{\infty}\arctan\left(\frac{\eta(b-a)}{\eta^2+(b-\lambda)(a-\lambda)}\right)d\mu(\lambda) \\
&+\frac{1}{\pi}\int_{a+\delta}^{b-\delta}\left[\pi-\arctan\left(\frac{b-\lambda}{\eta}\right)+\arctan\left(\frac{a-\lambda}{\eta}\right)\right]d\mu(\lambda) \\
&+\frac{1}{\pi}\int_{-\infty}^{a-\delta}\arctan\left(\frac{\eta(b-a)}{\eta^2+(b-\lambda)(a-\lambda)}\right)d\mu(\lambda).
\end{aligned}
$$

Note that in the two integrals with infinite limits, the arguments of arctan are positive. In the middle integral the variable λ runs between $a+\delta$ and $b-\delta$. For such λ, $\frac{b-\lambda}{\eta}\ge\frac{\delta}{\eta}$ and $\frac{a-\lambda}{\eta}\le-\frac{\delta}{\eta}$. So the right-hand side of the above inequality is dominated by

$$
\begin{aligned}
\frac{2\varepsilon}{5} &+ \frac{\eta}{\pi}\int_{b+\delta}^{\infty}\frac{b-a}{\eta^2+(b-\lambda)(a-\lambda)}d\mu(\lambda) \\
&+ \frac{\eta}{\pi}\int_{-\infty}^{a-\delta}\frac{b-a}{\eta^2+(b-\lambda)(a-\lambda)}d\mu(\lambda) \\
&+ \frac{1}{\pi}\int_{a+\delta}^{b-\delta}\left[\pi-2\arctan\frac{\delta}{\eta}\right]d\mu(\lambda).
\end{aligned}
$$

The first two integrals are finite (because of the properties of $d\mu$). The third one is dominated by $2(\frac{\pi}{2}-\arctan\frac{\delta}{\eta})[\mu(b)-\mu(a)]$. So we can choose η small enough to make each of the last three terms smaller than $\varepsilon/5$. This proves the theorem. ∎

We have shown above that all the terms occurring in the representation (V.42) are uniquely determined by the relations (V.43), (V.44), and (V.45).

Exercise V.4.13 *We have proved the relations (V.33), (V.36), (V.41) and (V.42) in that order. Show that all these are, in fact, equivalent. Hence, each of these representations is unique.*

Proposition V.4.14 *A Pick function φ is in the class $P(a,b)$ if and only if the measure μ associated with it in the representation (V.42) has zero mass on (a,b).*

Proof. Let $\varphi(x+i\eta) = u(x+i\eta) + iv(x+i\eta)$, where u, v are the real and imaginary parts of φ. If φ can be continued across (a,b), then as $\eta \downarrow 0$, on any closed subinterval $[c,d]$ of $(a,b), v(x+i\eta)$ converges uniformly to a bounded continuous function $v(x)$ on $[c,d]$. Hence,

$$\mu(d) - \mu(c) = \frac{1}{\pi}\int_c^d v(x)dx,$$

i.e., $d\mu(x) = \frac{1}{\pi}v(x)dx$. If the analytic continuation to the lower half-plane is by reflection across (a,b), then v is identically zero on $[c,d]$ and hence so is μ.

Conversely, if μ has no mass on (a,b), then for ζ in (a,b) the integral in (V.42) is convergent, and is real valued. This shows that the function φ can be continued from H_+ to H_- across (a,b) by reflection. ■

The reader should note that the above proposition shows that the converse of Theorem V.4.7 is also true.

It should be pointed out that the formula (V.42) defines two analytic functions, one on H_+ and the other on H_-. If these are denoted by φ and ψ, then $\varphi(\zeta) = \overline{\psi(\bar{\zeta})}$. So φ and ψ are reflections of each other. But they need not be analytic continuations of each other. For this to be the case, the measure μ should be zero on an interval (a,b) across which the function can be continued analytically.

Exercise V.4.15 *If a function f is operator monotone on the whole real line, then f must be of the form $f(t) = \alpha + \beta t$, $\alpha \in \mathbb{R}$, $\beta \geq 0$.*

Let us now look at a few simple examples.

Example V.4.16 *The function $\varphi(\zeta) = -\frac{1}{\zeta}$ is a Pick function. For this function, we see from (V.43) and (V.44) that $\alpha = \beta = 0$. Since φ is analytic everywhere in the plane except at 0, Proposition V.4.14 tells us that the measure μ is concentrated at the single point 0.*

Example V.4.17 *Let $\varphi(\zeta) = \zeta^{1/2}$ be the principal branch of the square root function. This is a Pick function. From (V.43) we see that*

$$\alpha = \text{Re}\,\varphi(i) = \text{Re}\,e^{i\pi/4} = \frac{1}{\sqrt{2}}.$$

From (V.44) we see that $\beta = 0$. If $\zeta = \lambda + i\eta$ is any complex number, then

$$\zeta^{1/2} = \left(\frac{|\zeta| + \lambda}{2}\right)^{1/2} + i\,\text{sgn}\,\eta \left(\frac{|\zeta| - \lambda}{2}\right)^{1/2},$$

where sgn η *is the sign of η, defined to be 1 if $\eta \geq 0$ and -1 if $\eta < 0$. So if $\eta \geq 0$, we have* Im $\varphi(\zeta) = \left(\frac{|\zeta|-\lambda}{2}\right)^{1/2}$. *As $\eta \downarrow 0$, $|\zeta|$ comes closer to $|\lambda|$. So,* Im $\varphi(\lambda + i\eta)$ *converges to 0 if $\lambda > 0$ and to $|\lambda|^{1/2}$ if $\lambda < 0$. Since φ is positive on the right half-axis, the measure μ has no mass at 0. The measure can now be determined from (V.45). We have, then*

$$\zeta^{1/2} = \frac{1}{\sqrt{2}} + \int_{-\infty}^{0} \left(\frac{1}{\lambda - \zeta} - \frac{\lambda}{\lambda^2 + 1}\right) \frac{|\lambda|^{1/2}}{\pi} d\lambda. \tag{V.46}$$

Example V.4.18 *Let $\varphi(\zeta) =$ Log ζ, where* Log *is the principal branch of the logarithm, defined everywhere except on $(-\infty, 0]$ by the formula* Log $\zeta = \ln|\zeta| + i$ Arg ζ. *The function* Arg ζ *is the principal branch of the argument, taking values in $(-\pi, \pi]$. We then have*

$$\begin{aligned} \alpha &= \text{Re}(\text{Log}\, i) = 0 \\ \beta &= \lim_{\eta \to \infty} \frac{\text{Log}(i\eta)}{i\eta} = 0. \end{aligned}$$

As $\eta \downarrow 0$, Im (Log$(\lambda + i\eta)$) *converges to π if $\lambda < 0$ and to 0 if $\lambda > 0$. So from (V.45) we see that, the measure μ is just the restriction of the Lebesgue measure to $(-\infty, 0]$. Thus,*

$$\text{Log}\,\zeta = \int_{-\infty}^{0} \left(\frac{1}{\lambda - \zeta} - \frac{\lambda}{\lambda^2 + 1}\right) d\lambda. \tag{V.47}$$

Exercise V.4.19 *For $0 < r < 1$, let ζ^r denote the principal branch of the function $\varphi(\zeta) = \zeta^r$. Show that*

$$\zeta^r = \cos\frac{r\pi}{2} + \frac{\sin r\pi}{\pi} \int_{-\infty}^{0} \left(\frac{1}{\lambda - \zeta} - \frac{\lambda}{\lambda^2 + 1}\right) |\lambda|^r d\lambda. \tag{V.48}$$

This includes (V.46) as a special case.

Let now f be any operator monotone function on $(0,\infty)$. We have seen above that f must have the form

$$f(t)=\alpha+\beta t+\int_{-\infty}^{0}\left(\frac{1}{\lambda-t}-\frac{\lambda}{\lambda^2+1}\right)d\mu(\lambda).$$

By a change of variables we can write this as

$$f(t)=\alpha+\beta t+\int_{0}^{\infty}\left(\frac{\lambda}{\lambda^2+1}-\frac{1}{\lambda+t}\right)d\mu(\lambda), \tag{V.49}$$

where $\alpha\in\mathbb{R}$, $\beta\geq 0$ and μ is a positive measure on $(0,\infty)$ such that

$$\int_{0}^{\infty}\frac{1}{\lambda^2+1}d\mu(\lambda)<\infty. \tag{V.50}$$

Suppose f is such that

$$f(0):=\lim_{t\to 0}f(t)>-\infty. \tag{V.51}$$

Then, it follows from (V.49) that μ must also satisfy the condition

$$\int_{0}^{1}\frac{1}{\lambda}d\mu(\lambda)<\infty. \tag{V.52}$$

We have from (V.49)

$$\begin{aligned} f(t)-f(0) &= \beta t+\int_{0}^{\infty}\left(\frac{1}{\lambda}-\frac{1}{\lambda+t}\right)d\mu(\lambda) \\ &= \beta t+\int_{0}^{\infty}\frac{t}{(\lambda+t)\lambda}\,d\mu(\lambda). \end{aligned}$$

Hence, we can write f in the form

$$f(t)=\gamma+\beta t+\int_{0}^{\infty}\frac{\lambda t}{\lambda+t}dw(\lambda), \tag{V.53}$$

where $\gamma=f(0)$ and $dw(\lambda)=\frac{1}{\lambda^2}d\mu(\lambda)$. From (V.50) and (V.52), we see that the measure w satisfies the conditions

$$\int_{0}^{\infty}\frac{\lambda^2}{\lambda^2+1}dw(\lambda)<\infty \text{ and } \int_{0}^{1}\lambda dw(\lambda)<\infty. \tag{V.54}$$

These two conditions can, equivalently, be expressed as a single condition

$$\int_0^\infty \frac{\lambda}{1+\lambda} dw(\lambda) < \infty. \tag{V.55}$$

We have thus shown that an operator monotone function on $(0, \infty)$ satisfying the condition (V.51) has a canonical representation (V.53), where $\gamma \in \mathbb{R}, \beta \geq 0$ and w is a positive measure satisfying (V.55).

The representation (V.53) is often useful for studying operator monotone functions on the positive half-line $[0, \infty)$.

Suppose that we are given a function f as in (V.53). If μ satisfies the conditions (V.54) then

$$\int_0^\infty \left(\frac{\lambda}{\lambda^2+1} - \frac{1}{\lambda} \right) \lambda^2 dw(\lambda) > -\infty,$$

and we can write

$$f(t) = \{\gamma - \int_0^\infty \left(\frac{\lambda}{\lambda^2+1} - \frac{1}{\lambda} \right) \lambda^2 dw(\lambda)\} + \beta t + \int_0^\infty \left(\frac{\lambda}{\lambda^2+1} - \frac{1}{\lambda+t} \right) \lambda^2 dw(\lambda).$$

So, if we put the number in braces above equal to α and $d\mu(\lambda) = \lambda^2 dw(\lambda)$, then we have a representation of f in the form (V.49).

Exercise V.4.20 *Use the considerations in the preceding paragraphs to show that, for* $0 < r \leq 1$ *and* $t > 0$, *we have*

$$t^r = \frac{\sin r\pi}{\pi} \int_0^\infty \frac{\lambda t}{\lambda + t} \lambda^{r-2} d\lambda. \tag{V.56}$$

(See Exercise V.1.10 also.)

Exercise V.4.21 *For* $t > 0$, *show that*

$$\log(1+t) = \int_1^\infty \frac{\lambda t}{\lambda + t} \lambda^{-2} d\lambda. \tag{V.57}$$

Appendix 1. Differentiability of Convex Functions

Let f be a real valued convex function defined on an interval I. Then f has some smoothness properties, which are listed below.

The function f is Lipschitz on any closed interval $[a, b]$ contained in I^0, the interior of I. So f is continuous on I^0.

At every point x in I^0, the right and left derivatives of f exist. These are defined, respectively, as

$$f'_+(x) := \lim_{y \downarrow x} \frac{f(y) - f(x)}{y - x},$$

$$f'_-(x) := \lim_{y \uparrow x} \frac{f(y) - f(x)}{y - x}.$$

Both these functions are monotonically increasing on I^0. Further,

$$\begin{aligned} \lim_{x \downarrow w} f'_\pm(x) &= f'_+(w), \\ \lim_{x \uparrow w} f'_\pm(x) &= f'_-(w). \end{aligned}$$

The function f is differentiable except on a countable set E in I^0, i.e., at every point x in $I^0 \backslash E$ the left and right derivatives of f are equal. Further, the derivative f' is continuous on $I^0 \backslash E$.

If a sequence of convex functions converges at every point of I, then the limit function is convex. The convergence is uniform on any closed interval $[a, b]$ contained in I^0.

Appendix 2. Regularisation of Functions

The convolution of two functions leads to a new function that inherits the stronger of the smoothness properties of the two original functions. This is the idea behind "regularisation" of functions.

Let φ be a real function of class C^∞ with the following properties: $\varphi \geq 0, \varphi$ is even, the support supp $\varphi = [-1, 1]$, and $\int \varphi = 1$. For each $\varepsilon > 0$, let $\varphi_\varepsilon(x) = \frac{1}{\varepsilon}\varphi(\frac{x}{\varepsilon})$. Then supp $\varphi_\varepsilon = [-\varepsilon, \varepsilon]$ and φ_ε has all the other properties of φ listed above. The functions φ_ε are called **mollifiers** or **smooth approximate identities**.

If f is a locally integrable function, we define its **regularisation of order** ε as the function

$$\begin{aligned} f_\varepsilon(x) = (f * \varphi_\varepsilon)(x) &:= \int f(x - y)\varphi_\varepsilon(y)dy \\ &= \int f(x - \varepsilon t)\varphi(t)dt. \end{aligned}$$

The family f_ε has the following properties.

1. Each f_ε is a C^∞ function.

2. If the support of f is contained in a compact set K, then the support of f_ε is contained in an ε-neighbourhood of K.

3. If f is continuous at x_0, then $\lim_{\varepsilon \downarrow 0} f_\varepsilon(x_0) = f(x_0)$.

4. If f has a discontinuity of the first kind at x_0, then $\lim_{\varepsilon \downarrow 0} f_\varepsilon(x_0) = 1/2\,[f(x_0+) + f(x_0-)]$. (A point x_0 is a point of discontinuity of the first kind if the left and right limits of f at x_0 exist; these limits are denoted as $f(x_0-)$ and $f(x_0+)$, respectively.)

5. If f is continuous, then $f_\varepsilon(x)$ converges to $f(x)$ as $\varepsilon \to 0$. The convergence is uniform on every compact set.

6. If f is differentiable, then, for every $\varepsilon > 0, (f_\varepsilon)' = (f')_\varepsilon$.

7. If f is monotone, then, as $\varepsilon \to 0, f_\varepsilon'(x)$ converges to $f'(x)$ at all points x where $f'(x)$ exists. (Recall that a monotone function can have discontinuities of the first kind only and is differentiable almost everywhere.)

V.5 Problems

Problem V.5.1. Show that the function $f(t) = \exp t$ is neither operator monotone nor operator convex on any interval.

Problem V.5.2. Let $f(t) = \frac{at+b}{ct+d}$, where a, b, c, d are real numbers such that $ad - bc > 0$. Show that f is operator monotone on every interval that does not contain the point $\frac{-d}{c}$.

Problem V.5.3. Show that the derivative of an operator convex function need not be operator monotone.

Problem V.5.4. Show that for $r < -1$, the function $f(t) = t^r$ on $(0, \infty)$ is not operator convex. (Hint: The function $f^{[1]}(1, t)$ cannot be continued analytically to a Pick function.) Together with the assertion in Exercise V.2.11, this shows that on the half-line $(0, \infty)$ the function $f(t) = t^r$ is operator convex if $-1 \le r \le 0$ or if $1 \le r \le 2$; and it is not operator convex for any other real r.

Problem V.5.5. A function g on $[0, \infty)$ is operator convex if and only if it is of the form

$$g(t) = \alpha + \beta t + \gamma t^2 + \int_0^\infty \frac{\lambda t^2}{\lambda + t}\, d\mu(\lambda),$$

where α, β are real numbers, $\gamma \ge 0$, and μ is a positive finite measure.

Problem V.5.6. Let f be an operator monotone function on $(0,\infty)$. Then $(-1)^{n-1}f^{(n)}(t) \geq 0$ for $n = 1, 2, \ldots$. [A function g on $(0,\infty)$ is said to be **completely monotone** if for all $n \geq 0$, $(-1)^n g^{(n)}(t) \geq 0$. There is a theorem of S.N. Bernstein that says that a function g is completely monotone if and only if there exists a positive measure μ such that $g(t) = \int\limits_0^\infty e^{-\lambda t} d\mu(\lambda)$.] The result of this problem says that the derivative of an operator monotone function on $(0,\infty)$ is completely monotone. Thus, f has a Taylor expansion $f(t) = \sum\limits_{n=0}^{\infty} a_n(t-1)^n$, in which the coefficients a_n are positive for all odd n and negative for all even n.

Problem V.5.7. Let f be a function mapping $(0,\infty)$ into itself. Let $g(t) = [f(t^{-1})]^{-1}$. Show that if f is operator monotone, then g is also operator monotone. If f is operator convex and $f(0) = 0$, then g is operator convex.

Problem V.5.8. Show that the function $f(\zeta) = -\cot\zeta$ is a Pick function. Show that in its canonical representation (V.42), $\alpha = \beta = 0$ and the measure μ is atomic with mass 1 at the points $n\pi$ for every integer n. Thus, we have the familiar series expansion

$$-\cot\zeta = \sum_{n=-\infty}^{\infty} \left[\frac{1}{n\pi - \zeta} - \frac{n\pi}{n^2\pi^2 + 1} \right].$$

Problem V.5.9. The aim of this problem is to show that if a Pick function φ satisfies the growth restriction

$$\sup_{\eta\to\infty} |\eta\varphi(i\eta)| < \infty, \tag{V.58}$$

then its representation (V.42) takes the simple form

$$\varphi(\zeta) = \int_{-\infty}^{\infty} \frac{1}{\lambda - \zeta} d\mu(\lambda), \tag{V.59}$$

where μ is a finite measure.

To see this, start with the representation (V.41). The condition (V.58) implies the existence of a constant M that bounds, for all $\eta > 0$, the quantity $\eta\varphi(i\eta)$, and hence also its real and imaginary parts. This gives two inequalities:

$$|\alpha\eta + \int\limits_{-\infty}^{\infty} \frac{\eta(1-\eta^2)\lambda}{\lambda^2+\eta^2} d\nu(\lambda)| \leq M,$$

$$|\beta\eta^2 + \eta^2 \int\limits_{-\infty}^{\infty} \frac{1+\lambda^2}{\lambda^2+\eta^2} d\nu(\lambda)| \leq M.$$

From the first, conclude that

$$\alpha = \lim_{\eta \to \infty} \int_{-\infty}^{\infty} \frac{(\eta^2 - 1)\lambda}{\lambda^2 + \eta^2} d\nu(\lambda) = \int_{-\infty}^{\infty} \lambda \, d\nu(\lambda).$$

From the second, conclude that $\beta = 0$ and

$$\int_{-\infty}^{\infty} \frac{\eta^2}{\lambda^2 + \eta^2} (1 + \lambda^2) d\nu(\lambda) \leq M.$$

Taking limits as $\eta \to \infty$, this gives

$$\int_{-\infty}^{\infty} (1 + \lambda^2) d\nu(\lambda) = \int_{-\infty}^{\infty} d\mu(\lambda) \leq M.$$

Thus, μ is a finite measure. From (V.41), we get

$$\varphi(\zeta) = \int_{-\infty}^{\infty} \lambda \, d\nu(\lambda) + \int_{-\infty}^{\infty} \frac{1 + \lambda\zeta}{\lambda - \zeta} d\nu(\lambda).$$

This is the same as (V.59).

Conversely, observe that if φ has a representation like (V.59), then it must satisfy the condition (V.58).

Problem V.5.10. Let f be a function on $(0, \infty)$ such that

$$f(t) = \alpha + \beta t - \int_{0}^{\infty} \frac{1}{\lambda + t} d\mu(\lambda),$$

where $\alpha \in \mathbb{R}$, $\beta \geq 0$ and μ is a positive measure such that $\int \frac{1}{\lambda} d\mu(\lambda) < \infty$. Then f is operator monotone. Find operator monotone functions that can not be expressed in this form.

V.6 Notes and References

Operator monotone functions were first studied in detail by K.Löwner (C. Loewner) in a seminal paper *Über monotone Matrixfunktionen*, Math. Z., 38 (1934) 177-216. In this paper, he established the connection between operator monotonicity, the positivity of the matrix of divided differences (Theorem V.3.4), and Pick functions. He also noted that the functions $f(t) = t^r$, $0 \leq r \leq 1$, and $f(t) = \log t$ are operator monotone on $(0, \infty)$.

Operator convex functions were studied, soon afterwards, by F. Kraus, *Über konvexe Matrixfunktionen*, Math. Z., 41(1936) 18-42.

In another well-known paper, *Beiträge zur Störungstheorie der Spectralzerlegung*, Math. Ann., 123 (1951) 415-438, E. Heinz used the theory of operator monotone functions to study several problems of perturbation theory for bounded and unbounded operators. The integral representation (V.41) in this context seems to have been first used by him. The operator monotonicity of the map $A \to A^r$ for $0 \le r \le 1$ is sometimes called the "Loewner-Heinz inequality", although it was discovered by Loewner.

J. Bendat and S. Sherman, *Monotone and convex operator functions*, Trans. Amer. Math. Soc., 79(1955) 58-71, provided a new perspective on the theorems of Loewner and Kraus. Theorem V.4.4 was first proved by them, and used to give a proof of Loewner's theorems.

A completely different and extremely elegant proof of Loewner's Theorem, based on the spectral theorem for (unbounded) selfadjoint operators was given by A. Korányi, *On a theorem of Löwner and its connections with resolvents of selfadjoint transformations*, Acta Sci. Math. Szeged, 17 (1956) 63-70.

Formulas like (V.13) and (V.22) were proved by Ju. L. Daleckii and S.G. Krein, *Formulas of differentiation according to a parameter of functions of Hermitian operators*, Dokl. Akad. Nauk SSSR, 76 (1951) 13-16. It was pointed out by M.G. Krein that the resulting Taylor formula could be used to derive conditions for operator monotonicity.

A concise presentation of the main ideas of operator monotonicity and convexity, including the approach of Daleckii and Krein, was given by C. Davis, *Notions generalizing convexity for functions defined on spaces of matrices*, in *Convexity: Proceedings of Symposia in Pure Mathematics*, American Mathematical Society, 1963, pp. 187-201. This paper also discussed other notions of convexity, examples and counterexamples, and was very influential.

A full book devoted to this topic is *Monotone Matrix Functions and Analytic Continuation*, by W.F. Donoghue, Springer-Verlag, 1974. Several ramifications of the theory and its connections with classical real and complex analysis are discussed here.

In a set of mimeographed lecture notes, *Topics on Operator Inequalities*, Hokkaido University, Sapporo, 1978, T. Ando provided a very concise modern survey of operator monotone and operator convex functions. Anyone who wishes to learn the Korányi method mentioned above should certainly read these notes.

A short proof of Löwner's Theorem appeared in G. Sparr, *A new proof of Löwner's theorem on monotone matrix functions*, Math. Scand., 47 (1980) 266-274.

In another brief and attractive paper, *Jensen's inequality for operators and Löwner's theorem*, Math. Ann., 258 (1982) 229-241, F. Hansen and G.K. Pedersen provided another approach.

Much of Sections 2, 3, and 4 are based on this paper of Hansen and Pedersen. For the latter parts of Section 4 we have followed Donoghue. We have also borrowed freely from Ando and from Davis. Our proof of Theorem V.1.9 is taken from M. Fujii and T. Furuta, *Löwner-Heinz, Cordes and Heinz-Kato inequalities*, Math. Japonica, 38 (1993) 73-78. Characterisations of operator convexity like the one in Exercise V.3.15 may be found in J.S. Aujla and H.L. Vasudeva, *Convex and monotone operator functions*, Ann. Polonici Math., 62 (1995) 1-11.

Operator monotone and operator convex functions are studied in R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Chapter 6. See also the interesting paper R.A. Horn, *The Hadamard product*, in C.R. Johnson, ed. *Matrix Theory and Applications*, American Mathematical Society, 1990.

A short, but interesting, section of the Marshall-Olkin book (cited in Chapter 2) is devoted to this topic. Especially interesting are some of the examples and connections with statistics that they give.

Among several applications of these ideas, there are two that we should mention here. Operator monotone functions arise often in the study of electrical networks. See, e.g., W.N. Anderson and G.E. Trapp, *A class of monotone operator functions related to electrical network theory*, Linear Algebra Appl., 15(1975) 53-67. They also occur in problems related to elementary particles. See, e.g., E. Wigner and J. von Neumann, *Significance of Löwner's theorem in the quantum theory of collisions*, Ann. of Math., 59 (1954) 418-433.

There are important notions of means of operators that are useful in the analysis of electrical networks and in quantum physics. An axiomatic approach to the study of these means was introduced by F. Kubo and T. Ando, *Means of positive linear operators*, Math. Ann., 249 (1980) 205-224. They establish a one-to-one correspondence between the class of operator monotone functions f on $[0, \infty)$ with $f(1) = 1$ and the class of operator means.

VI

Spectral Variation of Normal Matrices

Let A be an $n \times n$ Hermitian matrix, and let $\lambda_1^\downarrow(A) \geq \lambda_2^\downarrow(A) \geq \cdots \geq \lambda_n^\downarrow(A)$ be the eigenvalues of A arranged in decreasing order. In Chapter III we saw that $\lambda_j^\downarrow(A)$, $1 \leq j \leq n$, are continuous functions on the space of Hermitian matrices. This is a very special consequence of Weyl's Perturbation Theorem: if A, B are two Hermitian matrices, then

$$\max_j |\lambda_j^\downarrow(A) - \lambda_j^\downarrow(B)| \leq \|A - B\|.$$

In turn, this inequality is a special case of the inequality (IV.62), which says that if $\mathrm{Eig}^\downarrow(A)$ denotes the diagonal matrix with entries $\lambda_j^\downarrow(A)$ down its diagonal, then we have

$$|||\mathrm{Eig}^\downarrow(A) - \mathrm{Eig}^\downarrow(B)||| \leq |||A - B|||$$

for all Hermitian matrices A, B and for all unitarily invariant norms.

In this chapter we explore how far these results can be carried over to normal matrices. The first difficulty we face is that, if the matrices are not Hermitian, there is no natural way to order their eigenvalues. So, the problem has to be formulated in terms of optimal matchings. Even after this has been done, analogues of the inequalities above turn out to be a little more complicated. Though several good results are known, many await discovery.

VI.1 Continuity of Roots of Polynomials

Every polynomial of degree n with complex coefficients has n complex roots. These are unique, except for an ordering. It is thus natural to think of them as an *unordered* n-tuple of complex numbers. The space of such n-tuples is denoted by $\mathbb{C}^n_{sym}$. This is the quotient space obtained from the space $\mathbb{C}^n$ via the equivalence relation that identifies two n-tuples if their coordinates are permutations of each other. The space $\mathbb{C}^n_{sym}$ thus inherits a natural quotient topology from $\mathbb{C}^n$. It also has a natural metric: if $\lambda = \{\lambda_1, \ldots, \lambda_n\}$ and $\mu = \{\mu_1, \ldots, \mu_n\}$ are two points in $\mathbb{C}^n_{sym}$, then

$$d(\lambda, \mu) = \min_{\sigma} \max_{1 \le j \le n} |\lambda_j - \mu_{\sigma(j)}|,$$

where the minimum is taken over all permutations. See Problem II.5.9. This metric is called the **optimal matching distance** between λ and μ.

Exercise VI.1.1 *Show that the quotient topology on $\mathbb{C}^n_{sym}$ and the metric topology generated by the optimal matching distance are identical.*

Recall that, if

$$f(z) = z^n - a_1 z^{n-1} + a_2 z^{n-2} + \cdots + (-1)^n a_n \tag{VI.1}$$

is a monic polynomial with roots $\alpha_1, \ldots, \alpha_n$, then the coefficients a_j are elementary symmetric polynomials in the variables $\alpha_1, \ldots, \alpha_n$, i.e.,

$$a_j = \sum_{1 \le i_1 < \cdots < i_j \le n} \alpha_{i_1} \alpha_{i_2} \cdots \alpha_{i_j}. \tag{VI.2}$$

By the Fundamental Theorem of Algebra, we have a bijection $S : \mathbb{C}^n_{sym} \to \mathbb{C}^n$ defined as

$$S(\{\alpha_1, \ldots, \alpha_n\}) = (a_1, \ldots, a_n). \tag{VI.3}$$

Clearly S is continuous, by the definition of the quotient topology. We will show that S^{-1} is also continuous. For this we have to show that for every $\varepsilon > 0$, there exists $\delta > 0$ such that if $|a_j - b_j| < \delta$ for all j, then the optimal matching distance between the roots of the monic polynomials that have a_j and b_j as their coefficients is smaller than ε. Let $\xi_1, \ldots, \xi_k$ be the distinct roots of the monic polynomial f that has coefficients a_j. Given $\varepsilon > 0$, we can choose circles $\Gamma_j, 1 \le j \le k$, centred at ξ_j, each having radius smaller than ε and such that none of them intersects any other. Let Γ be the union of the boundaries of all these circles. Let $\eta = \inf_{z \in \Gamma} |f(z)|$. Then $\eta > 0$. Since Γ is a compact set, there exists a positive number δ such that if g is any monic polynomial with coefficients b_j, and $|a_j - b_j| < \delta$ for all j, then $|f(z) - g(z)| < \eta$ for all $z \in \Gamma$. So, by Rouché's Theorem f and g have the same number of zeroes inside each Γ_j, where the zeroes are counted with multiplicities. Thus we can pair each root of f with a root of g in

such a way that the distance between any two pairs is smaller than ε. In other words, the optimal matching distance between the roots of f and g is smaller than ε. We have thus proved the following.

Theorem VI.1.2 *The map S is a homeomorphism between $\mathbb{C}^n_{sym}$ and $\mathbb{C}^n$.*

The continuity of S^{-1} means that the *roots of a polynomial vary continuously with the coefficients.* Since the coefficients of its characteristic polynomial change continuously with a matrix, it follows that the eigenvalues of a matrix also vary continuously. More precisely, the map $\mathbf{M(n)} \to \mathbb{C}^n_{sym}$ that takes a matrix to the unordered tuple of its eigenvalues is continuous.

A different kind of continuity question is the following. If $z \to A(z)$ is a continuous map from a domain G in the complex plane into $\mathbf{M(n)}$, then do there exist n continuous functions $\lambda_1(z), \ldots, \lambda_n(z)$ on G such that for each z they are the eigenvalues of the matrix $A(z)$? The example below shows that this is not always the case.

Example VI.1.3 *Let $A(z) = \begin{pmatrix} 0 & z \\ 1 & 0 \end{pmatrix}$. The eigenvalues of $A(z)$ are $\pm z^{1/2}$. These cannot be represented by two single valued continuous functions on any domain G that contains zero.*

In two special situations, the answer to the question raised above is in the affirmative. If either the eigenvalues of $A(z)$ are all real, or if G is an interval on the real line, a continuous parametrisation of the eigenvalues of $A(z)$ is possible. This is shown below.

Consider the map from $\mathbb{R}^n_{sym}$ to $\mathbb{R}^n$ that rearranges an unordered n-tuple $\{\lambda_1, \ldots, \lambda_n\}$ in decreasing order as $(\lambda_1^\downarrow, \ldots, \lambda_n^\downarrow)$. From the majorisation relation (II.35) it follows that this map reduces distances, i.e.,

$$\max_{1 \le j \le n} |\lambda_j^\downarrow - \mu_j^\downarrow| \le d(\lambda, \mu).$$

Hence, in particular, this is a continuous map. So, if all the eigenvalues of $A(z)$ are real, enumerating them as $\lambda_1^\downarrow(z) \ge \cdots \ge \lambda_n^\downarrow(z)$ gives a continuous parametrisation for them. We should remark that while this is the most natural way of ordering real n-tuples, it is not always the most convenient. It could destroy the differentiability of these functions, which some other ordering might confer on them. For example, on any interval containing 0 the two functions $\pm t$ are differentiable. But rearrangement in the way above leads to the functions $\pm|t|$, which are not differentiable at 0.

For maps from an interval we have the following.

Theorem VI.1.4 *Let Λ be a continuous map from an interval I into the space $\mathbb{C}^n_{sym}$. Then there exist n continuous complex functions $\lambda_j(t)$ on I such that $\Lambda(t) = \{\lambda_1(t), \ldots, \lambda_n(t)\}$ for each $t \in I$.*

Proof. For brevity we will call n functions whose existence is asserted by the theorem a continuous selection for λ. Suppose a continuous selection $\lambda_j^{(1)}(t)$ exists on a subinterval I_1 and another continuous selection $\lambda_j^{(2)}(t)$ exists on a subinterval I_2. If I_1 and I_2 have a common point t_0, then $\{\lambda_j^{(1)}(t_0)\}$ and $\{\lambda_j^{(2)}(t_0)\}$ are identical up to a permutation. So a continuous selection exists on $I_1 \cup I_2$.

It follows that, if J is a subinterval of I such that each point of J has a neighbourhood on which a continuous selection exists, then a continuous selection exists on the entire interval J.

Now we can prove the theorem by induction on n. The statement is obviously true for $n = 1$. Suppose it is true for dimensions smaller than n. Let K be the set of all $t \in I$ for which all the n elements of $\Lambda(t)$ are equal. Then K is a closed subset of I. Let $L = I \backslash K$. Let $t_0 \in L$. Then $\Lambda(t_0)$ has at least two distinct elements. Collect all the copies of one of these elements. If these are k in number (i.e., k is the multiplicity of the chosen element), then the n elements of $\Lambda(t_0)$ are now divided into two groups with k and $n-k$ elements, respectively. These two groups have no element in common. Since $\Lambda(t)$ is continuous, for t sufficiently close to t_0 the elements of $\Lambda(t)$ also split into two groups of k and $n-k$ elements, each of which is continuous in t. By the induction hypothesis, each of these groups has a continuous selection in a neighbourhood of t_0. Taken together, they provide a continuous selection for Λ in this neighbourhood.

So, a continuous selection exists on each component of L. On its complement $K, \Lambda(t)$ consists of just one element $\lambda(t)$ repeated n times. Putting these together we obtain a continuous selection for $\Lambda(t)$ on all of I. ∎

Corollary VI.1.5 *Let $a_j(t), 1 \leq j \leq n$, be continuous complex valued functions defined on an interval I. Then there exist continuous functions $\alpha_1(t), \ldots, \alpha_n(t)$ that, for each $t \in I$, constitute the roots of the monic polynomial $z^n - a_1(t)z^{n-1} + \cdots + (-1)^n a_n(t)$.*

Corollary VI.1.6 *Let $t \to A(t)$ be a continuous map from an interval I into the space of $n \times n$ matrices. Then there exist continuous functions $\lambda_1(t), \ldots, \lambda_n(t)$ that, for each $t \in I$, are the eigenvalues of $A(t)$.*

VI.2 Hermitian and Skew-Hermitian Matrices

In this section we derive some bounds for the distance between the eigenvalues of a Hermitian matrix A and those of a skew-Hermitian matrix B. This will reveal several new facets of the general problem that are quite different from the case when both A, B are Hermitian.

Let us recall here, once again, the theorem that is the prototype of the results we seek.

Theorem VI.2.1 *(Weyl's Perturbation Theorem) Let A, B be Hermitian matrices with eigenvalues $\lambda_1^\downarrow(A) \geq \cdots \geq \lambda_n^\downarrow(A)$ and $\lambda_1^\downarrow(B) \geq \cdots \geq \lambda_n^\downarrow(B)$, respectively. Then*

$$\max_j |\lambda_j^\downarrow(A) - \lambda_j^\downarrow(B)| \leq \|A - B\|. \tag{VI.4}$$

We have seen two different proofs of this, one in Section III.2 and the other in Section IV.3. It is the latter idea which, in modified forms, will be used often in the following paragraphs.

Theorem VI.2.2 *Let A be a Hermitian and B a skew-Hermitian matrix. Let their eigenvalues $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ be arranged in such a way that*

$$|\alpha_1| \geq \cdots \geq |\alpha_n| \quad \textit{and} \quad |\beta_1| \geq \cdots \geq |\beta_n|. \tag{VI.5}$$

Then

$$\max_j |\alpha_j - \beta_{n-j+1}| \leq \|A - B\|. \tag{VI.6}$$

Proof. For a fixed index j, consider the eigenspaces of A and B corresponding to their eigenvalues $\{\alpha_1, \ldots, \alpha_j\}$ and $\{\beta_1, \ldots, \beta_{n-j+1}\}$, respectively. Let x be a unit vector in their intersection. Then

$$\begin{aligned}
\|A - B\|^2 &= \frac{1}{2}(\|A - B\|^2 + \|A + B\|^2) \\
&\geq \frac{1}{2}(\|(A - B)x\|^2 + \|(A + B)x\|^2) \\
&= \|Ax\|^2 + \|Bx\|^2 \\
&\geq |\alpha_j|^2 + |\beta_{n-j+1}|^2 = |\alpha_j - \beta_{n-j+1}|^2.
\end{aligned}$$

At the first step above, we used the equality $\|T\| = \|T^*\|$ valid for all T; at the third step we used the parallelogram law, and at the last step the fact that α_j is real and β_{n-j+1} is imaginary. ■

For Hermitian pairs A, B we have seen analogues of the inequality (VI.4) for other unitarily invariant norms. It is, therefore, natural to ask for similar kinds of results when A is Hermitian and B skew-Hermitian.

It is convenient to do this in the following setup. Let T be any matrix and let $T = A + iB$ be its Cartesian decomposition into real and imaginary parts, $A = \frac{T+T^*}{2}$ and $B = \frac{T-T^*}{2i}$. The theorem below gives majorisation relations between the eigenvalues of A and B, and the singular values of T. From these several inequalities can be obtained.

We will use the notation $\{x_j\}_j$ to mean an n-vector whose jth coordinate is x_j.

Theorem VI.2.3 *Let A, B be Hermitian matrices with eigenvalues α_j and β_j, respectively, ordered so that*

$$|\alpha_1| \geq \cdots \geq |\alpha_n| \quad \textit{and} \quad |\beta_1| \geq \cdots \geq |\beta_n|.$$

Let $T = A + iB$, and let s_j be the singular values of T. Then the following majorisation relations are satisfied:

$$\{|\alpha_j + i\beta_{n-j+1}|^2\}_j \prec \{s_j^2\}_j, \tag{VI.7}$$

$$\{1/2\ (s_j^2 + s_{n-j+1}^2)\}_j \prec \{|\alpha_j + i\beta_j|^2\}_j. \tag{VI.8}$$

Proof. For any two Hermitian matrices X, Y we have the majorisations (III.13):

$$\lambda^{\downarrow}(X) + \lambda^{\uparrow}(Y) \prec \lambda(X+Y) \prec \lambda^{\downarrow}(X) + \lambda^{\downarrow}(Y).$$

Choosing $X = A^2$, $Y = B^2$, this gives

$$\{|\alpha_j + i\beta_{n-j+1}|^2\}_j \prec \{s_j(A^2 + B^2)\}_j \prec \{|\alpha_j + i\beta_j|^2\}_j. \tag{VI.9}$$

Now note that

$$A^2 + B^2 = 1/2\ (T^*T + TT^*)$$

and

$$s_j(T^*T) = s_j(TT^*) = s_j^2.$$

So, choosing $X = \frac{T^*T}{2}$ and $Y = \frac{TT^*}{2}$ in the first majorisation above gives

$$1/2\ \{s_j^2 + s_{n-j+1}^2\}_j \prec \{s_j(A^2 + B^2)\}_j \prec \{s_j^2\}_j. \tag{VI.10}$$

Since majorisation is a transitive relation, the two assertions (VI.7) and (VI.8) follow from (VI.9) and (VI.10). ■

For each $p \geq 2$, the function $\varphi(t) = t^{p/2}$ is convex on $[0, \infty)$. So, by Corollary II.3.4, we obtain from (VI.7) and (VI.8) the weak majorisations

$$\{|\alpha_j + i\beta_{n-j+1}|^p\}_j \prec_w \{s_j^p\}_j, \tag{VI.11}$$

$$\frac{1}{2^{p/2}}\{(s_j^2 + s_{n-j+1}^2)^{p/2}\}_j \prec_w \{|\alpha_j + i\beta_j|^p\}_j. \tag{VI.12}$$

These two relations include the inequalities

$$\sum_{j=1}^{n} |\alpha_j + i\beta_{n-j+1}|^p \leq \sum_{j=1}^{n} s_j^p, \tag{VI.13}$$

$$\frac{1}{2^{p/2}} \sum_{j=1}^{n} (s_j^2 + s_{n-j+1}^2)^{p/2} \leq \sum_{j=1}^{n} |\alpha_j + i\beta_j|^p \tag{VI.14}$$

for $p \geq 2$.

If a_1 and a_2 are any two nonnegative real numbers, then the function $g(t) = (a_1^t + a_2^t)^{1/t}$ is monotonically decreasing on $0 < t < \infty$. So if $p \geq 2$, then

$$a_1^p + a_2^p \leq (a_1^2 + a_2^2)^{p/2}. \tag{VI.15}$$

Using this we get from (VI.14) the inequality

$$2^{1-p/2}\sum_{j=1}^{n} s_j^p \leq \sum_{j=1}^{n}|\alpha_j + i\beta_j|^p \tag{VI.16}$$

for $p \geq 2$.

Exercise VI.2.4 *For $0 < p \leq 2$, the function $\varphi(t) = t^{p/2}$ is concave on $[0,\infty)$. Use this to show that for these values of p, the weak majorisations (VI.11) and (VI.12) are valid with $\prec_w$ replaced by $\prec^w$. All the four inequalities (VI.13)-(VI.16) now go in the opposite direction.*

Let A be any matrix with eigenvalues $\alpha_1,\ldots,\alpha_n$, counted in any order. We have used the notation $\operatorname{Eig} A$ to mean a diagonal matrix that has entries α_j down its diagonal. If σ is a permutation, we will use the notation $\operatorname{Eig}_\sigma(A)$ for the diagonal matrix with entries $\alpha_{\sigma(1)},\ldots,\alpha_{\sigma(n)}$ down its diagonal. The symbol $\operatorname{Eig}^{|\downarrow|}(A)$ will mean the diagonal matrix whose diagonal entries are the eigenvalues of A in decreasing order of magnitude, i.e., the α_j arranged so that $|\alpha_1| \geq \cdots \geq |\alpha_n|$. In the same way, $\operatorname{Eig}^{|\uparrow|}(A)$ will stand for the diagonal matrix whose diagonal entries are the eigenvalues of A arranged in increasing order of magnitude, i.e., the α_j rearranged so that $|\alpha_1| \leq |\alpha_2| \leq \cdots \leq |\alpha_n|$.

With these notations, we have the following theorem for the distance between the eigenvalues of a Hermitian and a skew-Hermitian matrix, in the Schatten p-norms.

Theorem VI.2.5 *Let A be a Hermitian and B a skew-Hermitian matrix. Then,*

(i) for $2 \leq p \leq \infty$, we have

$$\|\operatorname{Eig}^{|\downarrow|}(A) - \operatorname{Eig}^{|\uparrow|}(B)\|_p \leq \|A - B\|_p, \tag{VI.17}$$

$$\|A - B\|_p \leq 2^{\frac{1}{2}-\frac{1}{p}}\|\operatorname{Eig}^{|\downarrow|}(A) - \operatorname{Eig}^{|\downarrow|}(B)\|_p; \tag{VI.18}$$

(ii) for $1 \leq p \leq 2$, we have

$$\|\operatorname{Eig}^{|\downarrow|}(A) - \operatorname{Eig}^{|\downarrow|}(B)\|_p \leq 2^{\frac{1}{p}-\frac{1}{2}}\|A - B\|_p, \tag{VI.19}$$

$$\|A - B\|_p \leq \|\operatorname{Eig}^{|\downarrow|}(A) - \operatorname{Eig}^{|\uparrow|}(B)\|_p. \tag{VI.20}$$

All the inequalities above are sharp. Further,

(iii) for $2 \leq p \leq \infty$, we have

$$\|\operatorname{Eig}^{|\downarrow|}(A)-\operatorname{Eig}^{|\uparrow|}(B)\|_p \leq \|\operatorname{Eig}(A)-\operatorname{Eig}_\sigma(B)\|_p \leq \|\operatorname{Eig}^{|\downarrow|}(A)-\operatorname{Eig}^{|\downarrow|}(B)\|_p \tag{VI.21}$$

for all permutations σ;

(iv) for $1 \leq p \leq 2$, we have

$$\|\operatorname{Eig}^{|\downarrow|}(A)-\operatorname{Eig}^{|\downarrow|}(B)\|_p \leq \|\operatorname{Eig}(A)-\operatorname{Eig}_\sigma(B)\|_p \leq \|\operatorname{Eig}^{|\downarrow|}(A)-\operatorname{Eig}^{|\uparrow|}(B)\|_p \tag{VI.22}$$

for all permutations σ.

Proof. For $p \geq 2$, the inequalities (VI.17) and (VI.18) follow immediately from (VI.13) and (VI.16). For $p = \infty$, the same inequalities remain valid by a limiting argument. The next two inequalities of the theorem follow from the fact that, for $1 \leq p \leq 2$, both of the inequalities (VI.13) and (VI.16) are reversed.

The special case of statements (i) and (ii) in which A and B commute is adequate for proving (iii) and (iv).

The sharpness of all the inequalities can be seen from the 2×2 example:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{VI.23}$$

Here $\|A - B\|_p = 2$ for all $1 \leq p \leq \infty$. The eigenvalues of A are ± 1, those of B are $\pm i$. Hence, for every permutation σ

$$\frac{\|\text{Eig}(A) - \text{Eig}_\sigma(B)\|_p}{\|A - B\|_p} = 2^{\frac{1}{p} - \frac{1}{2}}$$

for all $1 \leq p \leq \infty$. ■

Note that the inequality (VI.6) is included in (VI.17).

There are several features of these inequalities that are different from the corresponding inequality (IV.62) for a pair of Hermitian matrices A, B. First, the inequalities (VI.18) and (VI.19) involve a constant term on the right that is bigger than 1. Second, the best choice of this term depends on the norm $\|\cdot\|_p$. Third, the optimal matching of the eigenvalues of A with those of B – the one that will minimise the distance between them – changes with the norm. In fact, the best pairing for the norms $2 \leq p \leq \infty$ is the worst one for the norms $1 \leq p \leq 2$, and vice versa.

All these new features reveal that the spectral variation problem for pairs of normal matrices A, B is far more intricate than the one for Hermitian pairs.

Exercise VI.2.6 *Let A be a Hermitian and B a skew-Hermitian matrix. Show that for every unitarily invariant norm we have*

$$|||\text{Eig}^{|\downarrow|}(A) - \text{Eig}^{|\downarrow|}(B)||| \leq 2|||A - B|||, \tag{VI.24}$$

$$|||A - B||| \leq \sqrt{2}|||\text{Eig}^{|\downarrow|}(A) - \text{Eig}^{|\downarrow|}(B)|||. \tag{VI.25}$$

The term $\sqrt{2}$ in the second inequality cannot be replaced by anything smaller.

VI.3 Estimates in the Operator Norm

In this section we will obtain estimates of the distance between the eigenvalues of two normal matrices A and B in terms of $\|A - B\|$. Apart from

the optimal matching distance, which has already been introduced, we will consider other distances.

If L, M are two closed subsets of the complex plane $\mathbb{C}$, let

$$s(L,M) = \sup_{\lambda \in L} \operatorname{dist}(\lambda, M) = \sup_{\lambda \in L} \inf_{\mu \in M} |\lambda - \mu|. \tag{VI.26}$$

The **Hausdorff distance** between L and M is defined as

$$h(L,M) = \max(s(L,M),\ s(M,L)). \tag{VI.27}$$

Exercise VI.3.1 *Show that $s(L,M) = 0$ if and only if L is a subset of M. Show that the Hausdorff distance defines a metric on the collection of all closed subsets of $\mathbb{C}$.*

Note that $s(L,M)$ is the smallest number δ such that every element of L is within a distance δ of some element of M; and $h(L,M)$ is the smallest number δ for which this, as well as the symmetric assertion with L and M interchanged, is true.

Let $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_n\}$ be two unordered n-tuples of complex numbers. Let L and M be the *subsets* of $\mathbb{C}$ whose elements are the entries of these two tuples. If some entry among $\{\lambda_j\}$ or $\{\mu_j\}$ has multiplicity bigger than 1, then the cardinality of L or M is smaller than n.

Exercise VI.3.2 *(i) The Hausdorff distance $h(L,M)$ is always less than or equal to the optimal matching distance $d(\{\lambda_1, \ldots, \lambda_n\},\ \{\mu_1, \ldots, \mu_n\})$.*

(ii) When $n = 2$, the two distances are equal.

(iii) The triples $\{0, m - \varepsilon,\ m + \varepsilon\}$ and $\{m, \varepsilon, -\varepsilon\}$ provide an example in which $h(L,M) = \varepsilon$ and the optimal matching distance is $m - 2\varepsilon$. Thus, for $n \geq 3$, the second distance can be arbitrarily larger than the first.

If A is an $n \times n$ matrix, we will use the notation $\sigma(A)$ for both the subset of the complex plane that consists of all the eigenvalues of A, and for the unordered n-tuple whose entries are the eigenvalues of A counted with multiplicity. Since we will be talking of the distances $s(\sigma(A), \sigma(B))$, $h(\sigma(A), \sigma(B))$, and $d(\sigma(A), \sigma(B))$, it will be clear which of the two objects is being represented by $\sigma(A)$.

Note that the inequalities (VI.4) and (VI.6) say that

$$d(\sigma(A), \sigma(B)) \leq \|A - B\|, \tag{VI.28}$$

if either A and B are both Hermitian, or one is Hermitian and the other skew-Hermitian.

Theorem VI.3.3 *Let A be a normal and B an arbitrary matrix. Then*

$$s(\sigma(B), \sigma(A)) \leq \|A - B\|. \tag{VI.29}$$

Proof. Let $\varepsilon = \|A - B\|$. We have to show that if β is any eigenvalue of B, then β is within a distance ε of some eigenvalue α_j of A.

By applying a translation, we may assume that $\beta = 0$. If none of the α_j is within a distance ε of this, then A is invertible. Since A is normal, we have $\|A^{-1}\| = \frac{1}{\min|\alpha_j|} < \frac{1}{\varepsilon}$. Hence,

$$\|A^{-1}(B - A)\| \leq \|A^{-1}\| \, \|B - A\| < 1.$$

Since $B = A(I + A^{-1}(B - A))$, this shows that B is invertible. But then B could not have had a zero eigenvalue. ■

Another proof of this theorem goes as follows. Let A have the spectral resolution $A = \Sigma\alpha_j u_j u_j^*$, and let v be a unit vector such that $Bv = \beta v$. Then

$$\begin{aligned}\|A - B\|^2 &\geq \|(A - B)v\|^2 = \|\sum_j \alpha_j u_j^* v u_j - \beta \sum_j u_j^* v u_j\|^2 \\ &= \sum_j |\alpha_j - \beta|^2 |u_j^* v|^2.\end{aligned}$$

Since the u_j form an orthonormal basis, $\sum_j |u_j^* v|^2 = 1$. Hence, the above inequality can be satisfied only if $|\alpha_j - \beta|^2 \leq \|A - B\|^2$ for at least one index j.

Corollary VI.3.4 *If A and B are $n \times n$ normal matrices, then*

$$h(\sigma(A), \sigma(B)) \leq \|A - B\|. \tag{VI.30}$$

For $n = 2$, we have

$$d(\sigma(A), \sigma(B)) \leq \|A - B\|. \tag{VI.31}$$

This corollary also follows from the proposition below.

We will use the notation $D(a, \rho)$ for the open disk of radius ρ centred at a, and $\overline{D}(a, \rho)$ for the closure of this disk.

Proposition VI.3.5 *Let A and B be normal matrices, and let $\varepsilon = \|A - B\|$. If any disk $\overline{D}(a, \rho)$ contains k eigenvalues of A, then the disk $\overline{D}(a, \rho + \varepsilon)$ contains at least k eigenvalues of B.*

Proof. Without loss of generality, we may assume that $a = 0$. Suppose $\overline{D}(0, \rho)$ contains k eigenvalues of A but $\overline{D}(0, \rho + \varepsilon)$ contains less than k eigenvalues of B. Choose a unit vector x in the intersection of the eigenspace of A corresponding to its eigenvalues lying inside $\overline{D}(0, \rho)$ and the eigenspace of B corresponding to its eigenvalues lying outside $\overline{D}(0, \rho + \varepsilon)$. We then have $\|Ax\| \leq \rho$ and $\|Bx\| > \rho + \varepsilon$. We also have $\|Bx\| - \|Ax\| \leq \|(B - A)x\| \leq \|B - A\| = \varepsilon$. This is a contradiction. ■

Exercise VI.3.6 *Use the special case $\rho = 0$ of Proposition VI.3.5 to prove Corollary VI.3.4.*

Given a subset X of the complex plane and a matrix A, let $m_A(X)$ denote the number of eigenvalues of A inside X.

Exercise VI.3.7 *Let A, B be two $n \times n$ normal matrices. Let K_1, K_2 be two convex sets such that $m_A(K_1) \leq k$ and $m_B(K_2) \geq n - k + 1$. Then* $\operatorname{dist}(K_1, K_2) \leq \|A - B\|$. *[Hint: Let $\rho \to \infty$ in Proposition VI.3.5.]*

Exercise VI.3.8 *Use this to give another proof of Theorem VI.2.1.*

Exercise VI.3.9 *Let A, B be two $n \times n$ unitary matrices whose eigenvalues lie in a semicircle of the unit circle. Label both the sets of eigenvalues in the counterclockwise order. Then*

$$\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|.$$

Hence,

$$d(\sigma(A), \sigma(B)) \leq \|A - B\|.$$

Exercise VI.3.10 *Let T be the unit circle, I any closed arc in T, and for $\varepsilon > 0$ let I_ε be the arc $\{z \in T : |z - w| \leq \varepsilon$ for some $w \in I\}$. Let A, B be unitary matrices with $\|A - B\| = \varepsilon$. Show that $m_A(I) \leq m_B(I_\varepsilon)$.*

Theorem VI.3.11 *For any two unitary matrices,*

$$d(\sigma(A),\ \sigma(B)) \leq \|A - B\|.$$

Proof. The proof will use the Marriage Theorem (Theorem II.2.1) and the exercise above.

Let $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_n\}$ be the eigenvalues of A and B, respectively. Let Λ be any subset of $\{\lambda_1, \ldots, \lambda_n\}$. Let $\mu(\Lambda) = \{\mu_j : |\mu_j - \lambda_i| \leq \varepsilon$ for some $\lambda_i \in \Lambda\}$. By the Marriage Theorem, the assertion would be proved if we show that $|\mu(\Lambda)| \geq |\Lambda|$.

Let $I(\Lambda)$ be the set of all points on the unit circle T that are within distance ε of some point of Λ. Then $\mu(\Lambda)$ contains exactly those μ_j that lie in $I(\Lambda)$. Let $I(\Lambda)$ be written as a disjoint union of arcs $I_1, \ldots, I_r$. For each $1 \leq k \leq r$, let J_k be the arc contained in I_k all whose points are at distance $\geq \varepsilon$ from the boundary of I_k. Then $I_k = (J_k)_\varepsilon$.

From Exercise VI.3.10 we have

$$\sum_{k=1}^{r} m_A(J_k) \leq \sum_{k=1}^{r} m_B(I_k) = m_B(I(\Lambda)).$$

But, all the elements of Λ are in some J_k. This shows that $|\Lambda| \leq |\mu(\Lambda)|$. ■

There is one difference between Theorem VI.3.11 and most of our earlier results of this type. Now nothing is said about the order in which the

eigenvalues of A and B are arranged for the optimal matching. No canonical order can be prescribed in general. In Problem VI.8.3, we outline another proof of Theorem VI.3.11 which says, in effect, that for optimal matching the eigenvalues of A and B can be counted in the cyclic order on the circle *provided* the initial point is chosen properly. The catch is that this initial point depends on A and B and we do not know how to find it.

Exercise VI.3.12 *Let $A = c_1U_1$, $B = c_2U_2$, where U_1, U_2 are unitary matrices and c_1, c_2 are complex numbers. Show that $d(\sigma(A), \sigma(B)) \leq \|A - B\|$.*

By now we have seen that the inequality (VI.28) is valid in the following situations:

(i) A and B both Hermitian

(ii) A Hermitian and B skew-Hermitian

(iii) A and B both unitary (or both scalar multiples of unitaries)

(iv) A and B both 2×2 normal matrices.

The example below shows that this inequality breaks down for arbitrary normal matrices A, B when $n \geq 3$.

Example VI.3.13 *Let A be the 3×3 diagonal matrix with diagonal entries $\lambda_1 = 1$, $\lambda_2 = \frac{4+5\sqrt{3}\,i}{13}$, $\lambda_3 = \frac{-1+2\sqrt{3}\,i}{13}$. Let $v^T = \left(\sqrt{\frac{5}{8}}, \frac{1}{2}, \sqrt{\frac{1}{8}}\right)$ and let $U = I - 2vv^T$. Then U is a unitary matrix. Let $B = -U^*AU$. Then B is a normal matrix with eigenvalues $\mu_j = -\lambda_j$, $j = 1, 2, 3$. One can check that*

$$d(\sigma(A),\ \sigma(B)) = \sqrt{\frac{28}{13}},\quad \|A - B\| = \sqrt{\frac{27}{13}}.$$

So,

$$\frac{d(\sigma(A), \sigma(B))}{\|A - B\|} = 1.0183^+.$$

In the next chapter we will show that there exists a constant $c < 2.91$ such that for any two $n \times n$ normal matrices A, B

$$d(\sigma(A), \sigma(B)) \leq c\|A - B\|.$$

For Hermitian matrices A, B we have a reverse inequality:

$$\|A - B\| \leq \max_{1\leq j\leq n} |\lambda_j^{\downarrow}(A) - \lambda_j^{\uparrow}(B)|.$$

The quantity on the right is the distance between the eigenvalues of A and those of B when the "worst" pairing is made. An analogous result for normal matrices is proved below.

Theorem VI.3.14 *Let A and B be normal matrices with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_n\}$, respectively. Then, there exists a permutation σ such that*

$$\|A - B\| \leq \sqrt{2} \max_{1 \leq j \leq n} |\lambda_j - \mu_{\sigma(j)}|. \tag{VI.32}$$

Proof. The matrices $A \otimes I$ and $I \otimes B$ are both normal and commute with each other. Hence $A \otimes I - I \otimes B$ is normal. The eigenvalues of this matrix are all the differences $\lambda_i - \mu_j, 1 \leq i, j \leq n$. Hence

$$\|A \otimes I - I \otimes B\| = \max_{i,j} |\lambda_i - \mu_j|.$$

So, the inequality (VI.32) is equivalent to

$$\|A - B\| \leq \sqrt{2}\, \|A \otimes I - I \otimes B^T\|.$$

This is, in fact, true for all A, B and is proved below. ∎

Theorem VI.3.15 *For all matrices A, B*

$$\|A - B\| \leq \sqrt{2} \|A \otimes I - I \otimes B^T\|. \tag{VI.33}$$

Proof. We have to prove that for all x, y in $\mathbb{C}^n$

$$|\langle x, (A - B)y \rangle| \leq \sqrt{2} \|A \otimes I - I \otimes B^T\|\, \|x\|\, \|y\|.$$

We have

$$\begin{aligned} |\langle x, (A - B)y \rangle| &= |x^*Ay - x^*By| = |\text{tr}(Ayx^* - yx^*B)| \\ &\leq \|Ayx^* - yx^*B\|_1. \end{aligned}$$

The matrix $Ayx^* - yx^*B$ has rank at most 2. So,

$$\|Ayx^* - yx^*B\|_1 \leq \sqrt{2} \|Ayx^* - yx^*B\|_2.$$

Let $\bar{x}$ be the vector whose components are the complex conjugates of the components of x. Then with respect to the standard basis $e_i \otimes e_j$ of $\mathbb{C}^n \otimes \mathbb{C}^n$, the (i,j)-coordinate of the vector $(A \otimes I)(y \otimes \bar{x})$ is $\sum_k a_{ik} y_k \bar{x}_j$. This is also the (i,j)-entry of the matrix Ayx^*. In the same way, the (i,j)-entry of yx^*B is the (i,j)-coordinate of the vector $(I \otimes B^T)(y \otimes \bar{x})$. Thus, we have

$$\begin{aligned} \|Ayx^* - yx^*B\|_2 &= \|(A \otimes I - I \otimes B^T)(y \otimes \bar{x})\| \\ &\leq \|A \otimes I - I \otimes B^T\|\, \|y \otimes \bar{x}\| \\ &= \|A \otimes I - I \otimes B^T\|\, \|x\|\, \|y\|. \end{aligned}$$

This proves the theorem. ∎

The example (VI.23) shows that the inequality (VI.32) is sharp. Note that in this example A and B are both unitary. Also, A is Hermitian and B is skew-Hermitian. In contrast, the factor $\sqrt{2}$ in (VI.32) can be replaced by 1 if A, B are both Hermitian.

VI.4 Estimates in the Frobenius Norm

We will use the symbol S_n to mean the set of permutations on n symbols, as well as the set of $n \times n$ permutation matrices. (To every permutation σ there corresponds a unique matrix P that has entries 1 in the (i,j) place if and only if $j = \sigma(i)$, and all whose remaining entries are zero.) Let Ω_n be the set of all $n \times n$ doubly stochastic matrices. This is a convex polytope and by Birkhoff's Theorem (Theorem II.2.3) its extreme points are permutation matrices.

Theorem VI.4.1 *(Hoffman-Wielandt) Let A and B be normal matrices with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_n\}$, respectively. Then*

$$\min_{\sigma \in S_n} \left(\sum_{i=1}^{n} |\lambda_i - \mu_{\sigma(i)}|^2 \right)^{1/2} \leq \|A - B\|_2 \leq \max_{\sigma \in S_n} \left(\sum_{i=1}^{n} |\lambda_i - \mu_{\sigma(i)}|^2 \right)^{1/2}. \tag{VI.34}$$

Proof. Choose unitary matrices U, V such that $UAU^* = D_1$, $VBV^* = D_2$, where $D_1 = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and $D_2 = \text{diag}(\mu_1, \ldots, \mu_n)$. Then, by unitary invariance of the Frobenius norm, $\|A - B\|_2^2 = \|U^* D_1 U - V^* D_2 V\|_2^2 = \|D_1 W - W D_2\|^2$, where $W = UV^*$, another unitary matrix. If the matrix W has entries w_{ij}, this can be written as

$$\|A - B\|_2^2 = \sum_{i,j} |\lambda_i - \mu_j|^2 |w_{ij}|^2.$$

The matrix $(|w_{ij}|^2)$ is doubly stochastic. The map $(x_{ij}) \to \sum_{i,j} |\lambda_i - \mu_j|^2 x_{ij}$ is an affine function on the set Ω_n of doubly stochastic matrices. So it attains its minimum at one of the extreme points of Ω_n. Thus, there exists a permutation matrix (p_{ij}) such that

$$\|A - B\|_2^2 \geq \sum_{i,j} |\lambda_i - \mu_j|^2 p_{ij}.$$

If this matrix corresponds to the permutation σ, this says that

$$\|A - B\|_2^2 \geq \sum_{i} |\lambda_i - \mu_{\sigma(i)}|^2.$$

This proves the first inequality in (VI.34). The same argument for the maximum instead of the minimum gives the other inequality. ■

Note that for Hermitian matrices, the inequality (VI.34) was proved earlier in Problem III.6.15. In this case, we also proved that the same inequality is true for all unitarily invariant norms.

In general, there is no prescription for finding the permutation σ that minimises the Euclidean distance between the eigenvalues of A and those of B. However, if A is Hermitian with eigenvalues enumerated as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, then an enumeration of μ_j in which Re $\mu_1 \geq$ Re $\mu_2 \geq \cdots \geq$ Re μ_n is the best one. To see this, just note that if $\lambda_1 \geq \lambda_2$ and Re $\mu_1 \geq$ Re μ_2, then

$$|\lambda_1 - \mu_1|^2 + |\lambda_2 - \mu_2|^2 \leq |\lambda_1 - \mu_2|^2 + |\lambda_2 - \mu_1|^2.$$

The same argument shows that an enumeration for which the maximum distance is attained is one for which Re $\mu_1 \leq$ Re $\mu_2 \leq \cdots \leq$ Re μ_n. (What does this say when B is skew-Hermitian?)

Using the notations introduced in Section VI.2, the inequality (VI.34) can be rewritten as

$$\min_\sigma \|\text{Eig}(A) - \text{Eig}_\sigma(B)\|_2 \leq \|A - B\|_2 \leq \max_\sigma \|\text{Eig}(A) - \text{Eig}_\sigma(B)\|_2. \quad \text{(VI.35)}$$

There is another way of looking at this. Since the eigenvalues of a normal matrix completely determine the matrix up to a unitary conjugation, the inequality (VI.35) is equivalent to saying that for any two diagonal matrices A, B

$$\min_P \|A - PBP^*\|_2 \leq \|A - UBU^*\|_2 \leq \max_P \|A - PBP^*\|_2, \quad \text{(VI.36)}$$

where U is any unitary matrix and P varies over all permutation matrices.

Given any matrix B, let $\mathcal{U}_B$ be the set

$$\mathcal{U}_B = \{UBU^* : U \in \mathbf{U(n)}\},$$

where $\mathbf{U(n)}$ is the group consisting of unitary matrices. Then $\mathcal{U}_B$ is a compact set called the **unitary orbit** of B. For a fixed diagonal matrix A, consider the function $f(X) = \|A - X\|_2$. The inequality (VI.36) then says that if B is another diagonal matrix, then on the compact set $\mathcal{U}_B$ both the minimum and the maximum of f are attained at diagonal matrices (just some permutations of B). In other words, the minimum and the maximum on the unitary orbit are both contained in the **permutation orbit**.

This is an interesting fact from the point of view of calculus and geometry. We will see below that if A, B are real diagonal matrices, a stronger statement can be proved using calculus. This will also serve to introduce some elementary ideas of differential geometry used in later sections.

A differentiable function $U(t)$, where t is real and $U(t)$ is unitary, is called a **differentiable curve** through I if $U(0) = I$. Differentiating the equation $U(t)U(t)^* = I$ at $t = 0$ shows that for such a curve $U'(0)$ is skew-Hermitian. The matrix $U'(0)$ is called the **tangent vector** to $U(t)$ at I. If $K = U'(0)$, then e^{tK} is another differentiable curve through I with tangent vector K at I. Thus, the curves $U(t)$ and e^{tK} have the same tangent vector and so

represent the same curve *locally*, i.e., they are equal to the first degree of approximation. The **tangent space** to the manifold $\mathbf{U(n)}$ at the point I is the linear space that consists of all these tangent vectors. We have seen that this is the real vector space $\mathcal{K}(\mathbf{n})$ consisting of all skew-Hermitian matrices.

If $\mathcal{U}_A$ is the unitary orbit of a matrix A, then every differentiable curve through A can be represented locally as $e^{tK}Ae^{-tK}$ for some skew-Hermitian K. The derivative of this curve at $t=0$ is $KA-AK$. This is usually written as $[K,A]$ and called a **Lie bracket** or a **commutator**. Thus the tangent space to the manifold $\mathcal{U}_A$ at the point A is the space

$$T_A\mathcal{U}_A = \{[A,K] : K \in \mathcal{K}(\mathbf{n})\}. \tag{VI.37}$$

Note that this implies that $T_A\mathcal{U}_A$ is contained in $\mathcal{K}(\mathbf{n})$ if $A \in \mathcal{K}(\mathbf{n})$.

The sesquilinear form $\langle A, B\rangle = \operatorname{tr} A^*B$ is an inner product on the space $\mathbf{M(n)}$. The symbol $S^{\perp}$ will mean the orthogonal complement of a space S with respect to this inner product.

Lemma VI.4.2 *For every $A \in \mathcal{K}(\mathbf{n})$, the orthogonal complement of $T_A\mathcal{U}_A$ in $\mathcal{K}(\mathbf{n})$ is the set of all Y that commute with A.*

Proof. Let $Y \in \mathcal{K}(\mathbf{n})$. Then $Y \in (T_A\mathcal{U}_A)^{\perp}$ if and only if for every K in $\mathcal{K}(\mathbf{n})$ we have

$$\begin{aligned} 0 &= \langle Y, [A,K]\rangle = \operatorname{tr} Y^*(AK-KA) \\ &= -\operatorname{tr}(YAK - YKA) = \operatorname{tr}[A,Y]K. \end{aligned}$$

This is possible if and only if $[A,Y]=0$. ∎

The set of all matrices Y that commute with A is called the **commutant** or the **centraliser** of A, and is denoted as $Z(A)$. The lemma above says that in the space $\mathcal{K}(\mathbf{n})$, $(T_A\mathcal{U}_A)^{\perp} = Z(A)$ for every A.

Theorem VI.4.3 *Let $A \in \mathcal{K}(\mathbf{n})$ and let $f(X) = \|A-X\|_2$. Let B be any other element of $\mathcal{K}(\mathbf{n})$. Then B_0 is an extreme point for the function f on the unitary orbit $\mathcal{U}_B$ if and only if B_0 commutes with A.*

Proof. A point B_0 is an extreme point if and only if the straight line joining A and B_0 is perpendicular to $\mathcal{U}_B$ at B_0. By Lemma VI.4.2 this is so if and only if $A-B_0$ commutes with B_0, i.e., if and only if A commutes with B_0. ∎

For skew-Hermitian (or Hermitian) matrices A, B, this gives another proof of Theorem VI.4.1. However, in this case Theorem VI.4.3 says much more. From the first theorem we can conclude that if A and B are normal, then the *global* minimum and maximum of the (Frobenius) distance from A to $\mathcal{U}_B$ are attained among matrices that commute with A. The second

theorem says that when A and B are both Hermitian this is true for all *local* extrema as well.

This last statement is not true when A is Hermitian and B is skew-Hermitian. For in this case,

$$\|A - UBU^*\|_2^2 = \|A\|_2^2 + \|UBU^*\|_2^2 = \|A\|_2^2 + \|B\|_2^2$$

for all U. Thus the entire orbit $\mathcal{U}_B$ is at a constant distance from A. Hence, every point on $\mathcal{U}_B$ is an extremal point. However, not every point on $\mathcal{U}_B$ need commute with A.

VI.5 Geometry and Spectral Variation: the Operator Norm

The first theorem below says that if A is a normal matrix and B is any matrix close to A, then the optimal matching distance $d(\sigma(A), \sigma(B))$ is bounded by $\|A - B\|$. This is a *local* phenomenon; global versions of this are what we seek in the next paragraphs.

Theorem VI.5.1 *Let A be a normal matrix, and let B be any matrix such that $\|A - B\|$ is smaller than half the distance between any two distinct eigenvalues of A. Then $d(\sigma(A), \sigma(B)) \leq \|A - B\|$.*

Proof. Let $\alpha_1, \ldots, \alpha_k$ be all the distinct eigenvalues of A. Let $\varepsilon = \|A - B\|$. By Theorem VI.3.3, all the eigenvalues of B lie in the union of the disks $\overline{D}(\alpha_j, \varepsilon)$. By the hypothesis, these disks are mutually disjoint. We claim that if the eigenvalue α_j has multiplicity m_j, then the disk $\overline{D}(\alpha_j, \varepsilon)$ contains exactly m_j eigenvalues of B, counted with their respective multiplicities. Once this is established, the statement of the theorem is seen to follow easily.

Let $A(t) = (1-t)A + tB$, $0 \leq t \leq 1$. This is a continuous map from $[0,1]$ into the space of matrices; and we have $A(0) = A, A(1) = B$. Note that $\|A - A(t)\| = t\varepsilon$, and so all the eigenvalues of $A(t)$ also lie in the disks $\overline{D}(\alpha_j, \varepsilon)$ for each $0 \leq t \leq 1$. By Corollary VI.1.6, as t moves from 0 to 1, the eigenvalues of $A(t)$ trace continuous curves that join the eigenvalues of A to those of B. None of these curves can jump from one of the disks $\overline{D}(\alpha_j, \varepsilon)$ to another. So, if we start off with m_j such curves in the disk $\overline{D}(\alpha_j, \varepsilon)$, we must end up with exactly as many. ■

Example VI.3.13 shows that if no condition is imposed on B, then the conclusion of the theorem above is no longer valid, even when B is normal. However, this does suggest a new approach to the problem. Let A, B be normal matrices, and let $\gamma(t)$ be a curve joining A and B, such that each $\gamma(t)$ is a normal matrix. Then in a small neighbourhood of $\gamma(t)$ the spectral

variation inequality of Theorem VI.5.1 holds. So, the (total) spectral variation between the endpoints of the curve must be bounded by the length of this curve. This idea is made precise below.

Let **N** denote the set of normal matrices of a fixed size n. If A is an element of **N**, then so is tA for all real t. Thus the set **N** is path connected. However, **N** is not an affine set.

A continuous map γ from any interval $[a, b]$ into **N** will be called a **normal path** or a **normal curve**. If $\gamma(a) = A$ and $\gamma(b) = B$, we say that γ is a path joining A and B; A and B are then the endpoints of γ. The **length** of γ, with respect to the norm $\|\cdot\|$, is defined as

$$\ell_{\|\cdot\|}(\gamma) = \sup \sum_{k=0}^{m-1} \|\gamma(t_{k+1}) - \gamma(t_k)\|, \tag{VI.38}$$

where the supremum is taken over all partitions of $[a, b]$ as $a = t_0 < t_1 < \cdots < t_m = b$. If this length is finite, the path γ is said to be **rectifiable**. If the function γ is a piecewise C^1 function, then

$$\ell_{\|\cdot\|}(\gamma) = \int_a^b \|\gamma'(t)\| dt. \tag{VI.39}$$

Theorem VI.5.2 *Let A and B be normal matrices, and let γ be a rectifiable normal path joining them. Then*

$$d(\sigma(A),\ \sigma(B)) \leq \ell_{\|\cdot\|}(\gamma). \tag{VI.40}$$

Proof. For convenience, let us choose the parameter t to vary in $[0, 1]$. For $0 \leq r \leq 1$, let γ_r be that part of the curve which is parametrised by $[0, r]$. Let

$$G = \{r \in [0,1] : d(\sigma(A),\ \sigma(\gamma(r))) \leq \ell_{\|\cdot\|}(\gamma_r)\}.$$

The theorem will be proved if we show that the point 1 is in G.

Since the function γ, the arclength, and the distance d are all continuous in their arguments, the set G is closed. So it contains the point $g = \sup\ G$. We have to show that $g = 1$.

Suppose $g < 1$. Let $S = \gamma(g)$. Using Theorem VI.5.1, we can find a point t in $(g, 1]$ such that, if $T = \gamma(t)$, then $d(\sigma(S), \sigma(T)) \leq \|S - T\|$. But then

$$\begin{aligned} d(\sigma(A),\ \sigma(\gamma(t))) &\leq d(\sigma(A),\ \sigma(S)) + d(\sigma(S), \sigma(T)) \\ &\leq \ell_{\|\cdot\|}(\gamma_g) + \|S - T\| \\ &\leq \ell_{\|\cdot\|}(\gamma_t). \end{aligned}$$

By the definition of g, this is not possible. ■

An effective estimate of $d(\sigma(A), \sigma(B))$ can thus be obtained if one could find the length of the shortest normal path joining A and B. This is a

difficult problem since the geometry of the set **N** is poorly understood. However, the theorem above does have several interesting consequences.

Exercise VI.5.3 *Let $A, B \in \mathbf{N}$. Then the line segment joining A and B lies in $\mathbf{N}$ if and only if $A - B$ is in $\mathbf{N}$.*

Theorem VI.5.4 *If A, B are normal matrices such that $A - B$ is also normal, then $d(\sigma(A), \sigma(B)) \leq \|A - B\|$.*

Proof. The path γ consisting of the line segment joining A, B is a normal path by Exercise VI.5.3. Its length is $\|A - B\|$. ■

For Hermitian matrices A, B, the condition of the theorem above is satisfied. So this theorem includes Weyl's perturbation theorem as a special case.

A more substantial application of Theorem VI.5.2 is obtained as follows. It turns out that there exist normal matrices A, B for which $A - B$ is not normal, but there exists a normal path that joins them and has length $\|A - B\|$. Note that this path cannot be the line segment joining A and B; however, it has the same length as the line segment. What makes this possible is the fact that the metric under consideration is not Euclidean, and so geodesics need not always be straight lines. (Of course, by the definition of the arclength and the triangle inequality no path joining A, B could have length smaller than $\|A - B\|$.)

Let $\mathbf{S}$ be any subset of $\mathbf{M(n)}$. We will say that $\mathbf{S}$ is **metrically flat** in the metric induced by the norm $\|\cdot\|$ if any two points A, B of $\mathbf{S}$ can be joined by a path that lies entirely within $\mathbf{S}$ and has length $\|A - B\|$. To emphasize the dependence on the norm $\|\cdot\|$, we will also call such a set $\|\cdot\|$-flat.

Of course, every affine set is metrically flat. A nontrivial example of a $\|\cdot\|$-flat set is given by the theorem below. Let $\mathbf{U}$ be the set of $n \times n$ unitary matrices and $\mathbb{C} \cdot \mathbf{U}$ the set of all constant multiples of unitary matrices.

Theorem VI.5.5 *The set $\mathbb{C} \cdot \mathbf{U}$ is $\|\cdot\|$-flat.*

Proof. First note that $\mathbb{C} \cdot \mathbf{U}$ consists of just nonnegative real multiples of unitary matrices. Let $A_0 = r_0 U_0$ and $A_1 = r_1 U_1$ be any two elements of this set, where $r_0, r_1 \geq 0$. Choose an orthonormal basis in which the unitary matrix $U_1 U_0^{-1}$ is diagonal:

$$U_1 U_0^{-1} = \text{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n}),$$

where

$$|\theta_n| \leq \cdots \leq |\theta_1| \leq \pi.$$

Reduction to such a form can be achieved by a unitary conjugation. Such a process changes neither eigenvalues nor norms. So, we may assume that

all matrices are written with respect to the above orthonormal basis. Let

$$K = \operatorname{diag}(i\theta_1, \dots, i\theta_n).$$

Then, K is a skew-Hermitian matrix whose eigenvalues are in the interval $(-i\pi, i\pi]$. We have

$$\begin{aligned} \|A_0 - A_1\| &= \|r_0 I - r_1 U_1 U_0^{-1}\| = \max_j |r_0 - r_1 \exp(i\theta_j)| \\ &= |r_0 - r_1 \exp(i\theta_1)|. \end{aligned}$$

This last quantity is the length of the straight line joining the points r_0 and $r_1 \exp(i\theta_1)$ in the complex plane. Parametrise this line segment as $r(t)\exp(it\theta_1), 0 \le t \le 1$. This can be done except when $|\theta_1| = \pi$, an exceptional case to which we will return later. The equation above can then be written as

$$\begin{aligned} \|A_0 - A_1\| &= \int_0^1 |\, [r(t)\exp(it\theta_1)]' \,| dt \\ &= \int_0^1 |r'(t) + r(t) i\theta_1| dt. \end{aligned}$$

Now let $A(t) = r(t)\exp(tK)U_0, 0 \le t \le 1$. This is a smooth curve in $\mathbb{C} \cdot \mathbf{U}$ with endpoints A_0 and A_1. The length of this curve is

$$\begin{aligned} \int_0^1 \|A'(t)\| dt &= \int_0^1 \|r'(t)\exp(tK)U_0 + r(t)K\exp(tK)U_0\| dt \\ &= \int_0^1 \|r'(t)I + r(t)K\| dt, \end{aligned}$$

since $\exp(tK)U_0$ is a unitary matrix. But

$$\|r'(t)I + r(t)K\| = \max_j |r'(t) + ir(t)\theta_j| = |r'(t) + ir(t)\theta_1|.$$

Putting the last three equations together, we see that the path $A(t)$ joining A_0 and A_1 has length $\|A_0 - A_1\|$.

The exceptional case $|\theta_1| = \pi$ is much simpler. The piecewise linear path that joins A_0 to 0 and then to A_1 has length $r_0 + r_1$. This is equal to $|r_0 - r_1 \exp(i\theta_1)|$ and hence to $\|A_0 - A_1\|$. ■

Using Theorems VI.5.2 and VI.5.5, we obtain another proof of the result of Exercise VI.3.12.

Exercise VI.5.6 *Let A, B be normal matrices whose eigenvalues lie on concentric circles $C(A)$ and $C(B)$, respectively. Show that $d(\sigma(A), \sigma(B)) \leq \|A - B\|$.*

Theorem VI.5.7 *The set* **N** *consisting of $n \times n$ normal matrices is $\|\cdot\|$-flat if and only if $n \leq 2$.*

Proof. Let A, B be 2×2 normal matrices. If the eigenvalues of A and those of B lie on two parallel lines, we may assume that these lines are parallel to the real axis. Then the skew-Hermitian part of $A - B$ is a scalar, and hence $A - B$ is normal. The straight line joining A and B, then lies in **N**. If the eigenvalues do not lie on parallel lines, then they lie on two concentric circles. If α is the common centre of these circles, then A and B are in the set $\alpha + \mathbb{C} \cdot \mathbf{U}$. This set is $\|\cdot\|$-flat. Thus, in either case, A and B can be joined by a normal path of length $\|A - B\|$.

If $n \geq 3$, then **N** cannot be $\|\cdot\|$-flat because of Theorem VI.5.2 and Example VI.3.13. ■

Example VI.5.8 *Here is an example of a Hermitian matrix A and a skew-Hermitian matrix B that cannot be joined by a normal path of length $\|A - B\|$. Let*

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Then $\|A - B\| = 2$. If there were a normal path of length 2 joining A, B, then the midpoint of this path would be a normal matrix C such that $\|A - C\| = \|B - C\| = 1$. Since each entry of a matrix is dominated by its norm, this implies that $|c_{21} - 1| \leq 1$ and $|c_{21} + 1| \leq 1$. Hence $c_{21} = 0$. By the same argument, $c_{32} = 0$. So

$$A - C = \begin{pmatrix} * & * & * \\ 1 & * & * \\ * & 1 & * \end{pmatrix},$$

where $*$ represents an entry whose value is not yet known. But if $\|A - C\| = 1$, we must have

$$A - C = \begin{pmatrix} 0 & 0 & * \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Hence

$$C = \begin{pmatrix} 0 & 1 & * \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

But then C could not have been normal.

VI.6 Geometry and Spectral Variation: wui Norms

In this section we consider the possibility of extending to all (weakly) unitarily invariant norms, results obtained in the previous section for the operator norm. Given a wui norm τ, the **τ-optimal matching distance** between the eigenvalues of two (normal) matrices A, B is defined as

$$d_\tau(\sigma(A),\sigma(B)) = \min_P \tau(\operatorname{Eig} A - P(\operatorname{Eig} B)P^{-1}), \qquad \text{(VI.41)}$$

where, as before, $\operatorname{Eig} A$ is a diagonal matrix with eigenvalues of A down its diagonal (in any order) and where P varies over all permutation matrices. We want to compare this with the distance $\tau(A-B)$. The main result in this section is an extension of the path inequality in Theorem VI.5.2 to all wui norms. From this several interesting conclusions can be drawn.

Let us begin by an example that illustrates that not all results for the operator norm have straightforward extensions.

Example VI.6.1 *For $0 \le t \le \pi$, let $U(t) = \begin{pmatrix} 0 & 1 \\ e^{it} & 0 \end{pmatrix}$. Then,*

$$|||U(t) - U(0)||| = |1 - e^{it}| = 2 \sin \frac{t}{2},$$

for every unitarily invariant norm. In the trace norm (the Schatten 1-norm), we have

$$d_1(\sigma(U(t)),\ \sigma(U(0))) = 2|1 - e^{it/2}| = 4 \sin \frac{t}{4}.$$

So,

$$\frac{d_1(\sigma(U(t)),\sigma(U(0)))}{\|U(t) - U(0)\|_1} = \sec \frac{t}{4} > 1, \quad \text{for } t \neq 0.$$

Thus, we might have $d_1(\sigma(A),\sigma(B)) > \|A-B\|_1$, even for arbitrarily close normal matrices A, B. Compare this with Theorems VI.5.1 and VI.4.1.

The Q-norms are special in this respect, as we will see below.

Let Φ be any finite subset of $\mathbb{C}$. A map $F : \mathbb{C} \to \Phi$ is called a **retraction** onto Φ if $|z - F(z)| = \operatorname{dist}(z, \Phi)$, i.e., F maps every point in $\mathbb{C}$ to one of the points in Φ that is at the least distance from it. Such an F is not unique if Φ has more than one element.

Let Φ be a subset of $\mathbb{C}$ that has at most n elements, and let $\mathbf{N}(\Phi)$ be the set of all $n \times n$ normal matrices A such that $\sigma(A) \subset \Phi$. If F is a retraction onto Φ, then for every normal matrix B with eigenvalues $\{\beta_1, \ldots, \beta_n\}$ and for every A in $\mathbf{N}(\Phi)$ we have

$$\begin{aligned} \|B - F(B)\| &= \max_{1\le j\le n} |\beta_j - F(\beta_j)| = \max_j \operatorname{dist}(\beta_j, \Phi) \\ &= s(\sigma(B),\sigma(A)) \le \|B - A\| \end{aligned} \qquad \text{(VI.42)}$$

by Theorem VI.3.3. Note that the normality of B was required at the first step and that of A at the last. This inequality has a generalisation.

Theorem VI.6.2 *Let Φ be a finite set of cardinality at most n. Let F be a retraction onto Φ. Then for every normal matrix B and for every $A \in \mathbf{N}(\Phi)$ we have*

$$\|B - F(B)\|_Q \leq \|B - A\|_Q \tag{VI.43}$$

for every Q-norm.

Proof. By Exercise IV.2.10, the inequality (VI.43) is equivalent to the weak majorisation

$$[s(B - F(B))]^2 \prec_w [s(A - B)]^2.$$

If $\beta_1, \ldots, \beta_n$ are the eigenvalues of B, this is equivalent to saying that for all $1 \leq k \leq n$ we have

$$\sum_{j=1}^{k} |\beta_{i_j} - F(\beta_{i_j})|^2 \leq \sum_{j=1}^{k} s_j^2(A - B)$$

for every choice of indices $i_1, \ldots, i_k$.

By Ky Fan's maximum principle (Exercise II.1.13)

$$\sum_{j=1}^{k} s_j^2(A - B) = \max \sum_{j=1}^{k} \|(A - B)v_j\|^2,$$

where the maximum is taken over all orthonormal k-tuples $v_1, \ldots, v_k$. In particular, if e_j are unit vectors such that $Be_j = \beta_j e_j$, then

$$\sum_{j=1}^{k} s_j^2(A - B) \geq \sum_{j=1}^{k} \|(A - \beta_{i_j})e_{i_j}\|^2.$$

But if β is any complex number and e any unit vector, then $\|(A - \beta)e\| \geq \text{dist}(\beta, \sigma(A))$. (See the second proof of Theorem VI.3.3.) Hence, we have

$$\sum_{j=1}^{k} s_j^2(A - B) \geq \sum_{j=1}^{k} |\beta_{i_j} - F(\beta_{i_j})|^2,$$

and this completes the proof. ∎

Exercise VI.6.3 *Show that the assertion of Theorem VI.6.2 is not true for the Schatten p-norms, $1 \leq p < 2$. (See the example in (VI.23).)*

Corollary VI.6.4 *Let A be a normal matrix, and let B be another normal matrix such that $\|A - B\|$ is smaller than half the distance between the distinct eigenvalues of A. Then*

$$d_Q(\sigma(A), \sigma(B)) \leq \|A - B\|_Q$$

for every Q-norm. (The quantity on the left is the Q-norm optimal matching distance.)

Proof. Let $\varepsilon = \|A - B\|$. In the proof of Theorem VI.5.1 we saw that all the eigenvalues of B lie within the region comprising of disks of radius ε around the eigenvalues of A. Further, each such disk contains as many eigenvalues of A as of B (multiplicities counted). The retraction F of Theorem VI.6.2 then achieves a one-to-one pairing of the eigenvalues of A and those of B. ∎

Replacing the operator norm by any other norm τ in (VI.38), we can define the τ-length of a path γ by the same formula. Denote this by $\ell_\tau(\gamma)$.

Exercise VI.6.5 *Let A and B be normal matrices, and let γ be a normal path joining them. Then for every Q-norm we have*

$$d_Q(\sigma(A), \sigma(B)) \leq \ell_Q(\gamma).$$

This includes Theorem VI.5.2 as a special case.

We will now extend this inequality to its broadest context.

Proposition VI.6.6 *Let A be a normal matrix and let δ be half the minimum distance between distinct eigenvalues of A. Then there exists a positive number M (depending on δ and the dimension n) such that any normal matrix B with $\|A - B\| \leq \delta$ has a representation $B = UB'U^*$, where B' commutes with A and U is a unitary matrix with $\|I - U\| \leq M\|A - B\|$.*

Proof. Let $\alpha_j, 1 \leq j \leq r$, be the distinct eigenvalues of A, and let m_j be the multiplicity of α_j. Choose an orthonormal basis in which $A = \oplus_j \alpha_j I_j$, where $I_j, 1 \leq j \leq r$, are identity submatrices of dimensions m_j. By the argument used in the proof of Theorem VI.5.1, the eigenvalues of B can be grouped into diagonal blocks D_j, where D_j has dimension m_j and every eigenvalue of D_j is within distance δ of α_j. This implies that

$$\|(\alpha_j I_k - D_k)^{-1}\| \leq \frac{1}{\delta} \quad \text{if} \quad j \neq k.$$

If $D = \oplus_j D_j$, then there exists a unitary matrix W such that $B = WDW^*$.

With respect to the above splitting, let W have the block decomposition $W = [W_{jk}], 1 \leq j, k \leq r$. Then

$$\begin{aligned} \|A - B\| &= \|A - WDW^*\| = \|AW - WD\| \\ &= \|[W_{jk}(\alpha_j I_k - D_k)]\|. \end{aligned}$$

Hence, for $j \neq k$,

$$\|W_{jk}\| \leq \|(\alpha_j I_k - D_k)^{-1}\| \; \|A - B\| \leq \frac{1}{\delta}\|A - B\|.$$

Hence, there exists a constant K that depends only on δ and n such that

$$\|W - \oplus_j W_{jj}\| \leq K\|A - B\|.$$

Let $X = \oplus_j W_{jj}$. This is the diagonal part in the block decomposition of W. Hence, $\|X\| \leq 1$ by the pinching inequality. Let $W_{jj} = V_j P_j$ be the polar decomposition of W_{jj} with V_j unitary and P_j positive. Then

$$\|W_{jj} - V_j\| = \|P_j - I_j\| \leq \|P_j^2 - I_j\|,$$

since P_j is a contraction. Let $V = \oplus_j V_j$. Then V is unitary and from the above inequality, we see that

$$\|X - V\| \leq \|X^*X - I\| = \|X^*X - W^*W\|.$$

Hence,

$$\begin{aligned} \|W - V\| &\leq \|W - X\| + \|X - V\| \leq \|W - X\| + \|X^*X - W^*W\| \\ &\leq \|W - X\| + \|(X^* - W^*)X\| + \|W^*(X - W)\| \\ &\leq 3\|W - X\| \leq 3K\|A - B\|. \end{aligned}$$

If we put $U = WV^*$ and $M = 3K$, we have $\|I - U\| \leq M\|A - B\|$ and $B = WDW^* = UVDV^*U^* = UB'U^*$, where $B' = VDV^*$. Since B' is block-diagonal with diagonal blocks of size m_j, it commutes with A. This completes the proof. ■

Proposition VI.6.7 *Given a normal matrix A, a wui norm τ and an $\varepsilon > 0$, there exists a small neighbourhood of A such that for any normal matrix B in this neighbourhood we have*

$$d_\tau(\sigma(A), \sigma(B)) \leq (1 + \varepsilon)\tau(A - B).$$

Proof. Choose B so close to A that the conditions of Proposition VI.6.6 are satisfied. Let U, B', M be as in that proposition.

Let $S = U - I$, so that $U = I + S$ and $U^* = I - S + S^2U^*$. Then

$$A - B = A - B' + [B', S] + UB'SU^* - B'S.$$

Hence,

$$\tau(A - B' + [B', S]) \leq \tau(A - B) + \tau(UB'SU^* - B'S).$$

Since A and B' are commuting normal matrices, they can be diagonalised simultaneously by a unitary conjugation. In this basis the diagonal of $[B', S]$ will be zero. So, by the pinching inequality

$$\tau(A - B') \leq \tau(A - B' + [B', S]).$$

The two inequalities above give

$$d_\tau(\sigma(A), \sigma(B)) \leq \tau(A - B) + \tau(UB'SU^* - B'S).$$

Now choose k such that

$$\frac{1}{k} \leq \frac{\tau(X)}{\|X\|} \leq k \text{ for all } X.$$

Then, using Proposition VI.6.6, we get

$$\begin{aligned} \tau(UB'SU^* - B'S) &\leq 2\tau(B'S) \leq 2kM\|B\| \, \|A - B\| \\ &\leq 2k^2 M\|B\|\tau(A - B). \end{aligned}$$

Now, if B is so close to A that we have $2k^2 M\|B\| \leq \varepsilon$, then the inequality of the proposition is valid. ■

Theorem VI.6.8 *Let A, B be normal matrices, and let γ be any normal path joining them. Then there exists a permutation matrix P such that for every wui norm τ we have*

$$\tau(\text{Eig } A - P(\text{Eig } B)P^{-1}) \leq \ell_\tau(\gamma). \qquad \text{(VI.44)}$$

Proof. For convenience, let $\gamma(t)$ be parametrised on the interval $[0, 1]$. Let $\gamma(0) = A$, $\gamma(1) = B$. By Theorem VI.1.4, there exist continuous functions $\lambda_1(t), \ldots, \lambda_n(t)$ that represent the eigenvalues of the matrix $\gamma(t)$ for each t. Let $D(t)$ be the diagonal matrix with diagonal entries $\lambda_j(t)$. We will show that

$$\tau(D(0) - D(1)) \leq \ell_\tau(\gamma). \qquad \text{(VI.45)}$$

Let τ be any wui norm, and let ε be any positive number. Let $\gamma[s, t]$ denote the part of the path $\gamma(\cdot)$ that is defined on $[s, t]$. Let

$$G = \{t : \tau(D(0) - D(t)) \leq (1 + \varepsilon)\ell_\tau(\gamma[0, t])\}. \qquad \text{(VI.46)}$$

Because of continuity, G is a closed set and hence it includes its supremum g. We will show that $g = 1$. If this is not the case, then we can choose $g' > g$ so close to g that Proposition VI.6.7 guarantees

$$\tau(D(g) - PD(g')P^{-1}) \leq (1 + \varepsilon)\tau(\gamma(g) - \gamma(g')), \qquad \text{(VI.47)}$$

for some permutation matrix P. Now note that

$$\tau(D(g) - P^{-1}D(g)P) \leq \tau(D(g') - D(g)) + \tau(D(g) - PD(g')P^{-1}),$$

and hence if g' is sufficiently close to g, we will have $\tau(D(g) - P^{-1}D(g)P)$ small relative to the minimum distance between the distinct eigenvalues of $D(g)$. We thus have $D(g) = P^{-1}D(g)P$. Hence

$$\tau(D(g) - D(g')) = \tau(P^{-1}D(g)P - D(g')) = \tau(D(g) - PD(g')P^{-1}).$$

So, from (VI.47),

$$\tau(D(g) - D(g')) \leq (1+\varepsilon)\tau(\gamma(g) - \gamma(g')).$$

From the definition of g as the supremum of the set G in (VI.46), we have

$$\tau(D(0) - D(g)) \leq (1+\varepsilon)\ell_\tau(\gamma[0,g]).$$

Combining the two inequalities above, we get

$$\tau(D(0) - D(g')) \leq (1+\varepsilon)\ell_\tau(\gamma[0,g']).$$

This contradicts the definition of g. So $g = 1$. ■

The inequality (VI.45) tells us not only that for all normal A, B and for all wui norms τ we have

$$d_\tau(\sigma(A), \sigma(B)) \leq \ell_\tau(\gamma), \tag{VI.48}$$

but also that a matching of $\sigma(A)$ with $\sigma(B)$ can be chosen which makes this work simultaneously for all τ. Further, this matching is the natural one obtained by following the curves $\lambda_j(t)$ that describe the eigenvalues of the family $\gamma(t)$.

Several corollaries can be obtained now.

Theorem VI.6.9 *Let A, B be unitary matrices, and let K be any skew-Hermitian matrix such that $BA^{-1} = \exp K$. Then, for every unitarily invariant norm $|||\cdot|||$, we have,*

$$d_{|||\cdot|||}(\sigma(A), \sigma(B)) \leq |||K|||. \tag{VI.49}$$

Proof. Let $\gamma(t) = (\exp tK)A$, $0 \leq t \leq 1$. Then $\gamma(t)$ is unitary for all $t, \gamma(0) = A, \gamma(1) = B$. So, by Theorem VI.6.8,

$$d_{|||\cdot|||}(\sigma(A), \sigma(B)) \leq \int_0^1 |||\gamma'(t)|||dt.$$

But $\gamma'(t) = K(\exp tK)A$. So $|||\gamma'(t)||| = |||K|||$. ■

Theorem VI.6.10 *Let A, B be unitary matrices. Then for every unitarily invariant norm*

$$d_{|||\cdot|||}(\sigma(A), \sigma(B)) \leq \frac{\pi}{2}|||A - B|||. \tag{VI.50}$$

Proof. In view of Theorem VI.6.9, we need to show that

$$\inf\{|||K||| : BA^{-1} = \exp K\} \leq \frac{\pi}{2}|||A - B|||.$$

Choose a K whose eigenvalues are contained in the interval $(-i\pi, i\pi]$. By applying a unitary conjugation, we may assume that $K = \text{diag}(i\theta_1, \ldots, i\theta_n)$. Then

$$|||A - B||| = |||I - BA^{-1}||| = |||\text{diag}(1 - e^{i\theta_1}, \ldots, 1 - e^{i\theta_n})|||.$$

But if $-\pi < \theta \leq \pi$, then $|\theta| \leq \frac{\pi}{2}|1 - e^{i\theta}|$. Hence, $|||K||| \leq \frac{\pi}{2}|||A - B|||$ for every unitarily invariant norm. ■

We now give an example to show that the factor $\pi/2$ in the inequality (VI.50) cannot be reduced if the inequality is to hold for all unitarily invariant norms and all dimensions. Recall that for the operator norm and for the Frobenius norm we have the stronger inequality with 1 instead of $\pi/2$ (Theorem VI.3.11 and Theorem VI.4.1).

Example VI.6.11 *Let A_+ and A_- be the unitary matrices obtained by adding an entry ± 1 in the bottom left corner to an upper Jordan matrix, i.e.,*

$$A_{\pm} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ . & . & . & \cdots & . \\ 0 & 0 & 0 & \cdots & 1 \\ \pm 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Then for the trace norm we have $\|A_+ - A_-\|_1 = 2$. The eigenvalues of $A_{\pm}$ are the n roots of ± 1. One can see that the $\|\cdot\|_1$-optimal matching distance between these two n-tuples approaches π as $n \to \infty$.

The next theorem is a generalisation of, and can be proved using the same idea as, Theorem VI.5.4.

Theorem VI.6.12 *If A, B are normal matrices such that $A - B$ is also normal, then for every wui norm τ*

$$d_\tau(\sigma(A), \sigma(B)) \leq \tau(A - B). \tag{VI.51}$$

This inequality, or rather just its special case when τ is restricted to unitarily invariant norms and A, B are Hermitian, can be used to get yet another proof of Lidskii's Theorem. We have seen this argument earlier in Chapter IV. The stronger result we now have at our disposal gives a stronger version of Lidskii's Theorem. This is shown below.

Let x, y be elements of $\mathbb{C}^n$. We will say that x is **majorised** by y, in symbols $x \prec y$, if x is a convex combination of vectors obtained from y by permuting its coordinates, i.e., $x = \Sigma a_\sigma y_\sigma$, a finite sum in which each y_σ is

a vector whose coordinates are obtained by applying the permutation σ to the coordinates of y and a_σ are positive numbers with $\Sigma a_\sigma = 1$. When x, y are real vectors, this is already familiar. We will say x is **softly majorised** by y, in symbols $x \prec_s y$, if we can write x as a finite sum $x = \Sigma z_\sigma y_\sigma$ in which z_σ are complex numbers such that $\Sigma |z_\sigma| \leq 1$.

Exercise VI.6.13 *Let x, y be two vectors in $\mathbb{C}^n$ such that $\sum_{j=1}^{n} x_j$ and $\sum_{j=1}^{n} y_j$ are not zero. If $x \prec_s y$ and $\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} y_j$, then $x \prec y$.*

Proposition VI.6.14 *Let A, B be $n \times n$ normal matrices and let $\lambda(A)$, $\lambda(B)$ be two n-vectors whose coordinates are the eigenvalues of A, B, respectively. Then $\tau(A) \leq \tau(B)$ for all wui norms τ if and only if $\lambda(A) \prec_s \lambda(B)$.*

Proof. Suppose $\tau(A) \leq \tau(B)$ for all wui norms τ. Then, using Theorem IV.4.7, we can write the diagonal matrix $\text{Eig}(A)$ as a finite sum $\text{Eig}(A) = \Sigma z_k U_k \text{Eig}(B) U_k^*$, in which U_k are unitary matrices and $\Sigma |z_k| \leq 1$. This shows that $\lambda(A) = \Sigma z_k S_k(\lambda(B))$, where each S_k is an orthostochastic matrix. (An orthostochastic matrix S is a doubly stochastic matrix such that $s_{ij} = |u_{ij}|^2$, where u_{ij} are the entries of a unitary matrix.) By Birkhoff's Theorem each S_k is a convex combination of permutation matrices. Hence, $\lambda(A) \prec_s \lambda(B)$. The converse follows by the same argument without recourse to Birkhoff's Theorem. ■

Theorem VI.6.15 *Let A, B be normal matrices such that $A - B$ is also normal. Then the eigenvalues of A and B can be arranged in such a way that if $\lambda(A)$ and $\lambda(B)$ are the n-vectors with these eigenvalues as their coordinates, then*

$$\lambda(A) - \lambda(B) \prec \lambda(A - B). \tag{VI.52}$$

Proof. Use Theorem VI.6.8 and the observation in Theorem VI.6.12 to conclude that we can arrange the eigenvalues in such a way that

$$\tau(\text{Eig}\, A - \text{Eig}\, B) \leq \tau(A - B)$$

for every wui norm τ. By Proposition VI.6.14, this is equivalent to saying

$$\lambda(A) - \lambda(B) \prec_s \lambda(A - B),$$

where $\lambda(A)$ is the vector whose entries are the diagonal entries of the diagonal matrix $\text{Eig}\, A$. By a small perturbation, if necessary, we may assume that $\text{tr} A \neq \text{tr}\, B$. Since the components of the vectors $\lambda(A) - \lambda(B)$ and $\lambda(A - B)$ must have the same (nonzero) sum, we have in fact majorisation rather than just the soft majorisation proved above. ■

We can call this the **Lidskii Theorem for normal matrices**. It includes the classical Lidskii Theorem as a special case.

Exercise VI.6.16 *Let A, B be normal matrices such that $A - B$ is also normal. Let τ be any wui norm. Show that there exists a permutation matrix P such that*

$$\tau(A - B) \leq \tau(\text{Eig}\, A - P(\text{Eig}\, B)P^{-1}). \qquad \text{(VI.53)}$$

VI.7 Some Inequalities for the Determinant

The determinant of the sum $A + B$ of two matrices has no simple relation with the determinants of A and B. Some interesting inequalities can be derived using ideas introduced in this chapter. These are proved below.

Theorem VI.7.1 *Let A and B be Hermitian matrices with eigenvalues $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$, respectively. Then*

$$\min_\sigma \prod_{i=1}^n (\alpha_i + \beta_{\sigma(i)}) \leq \det(A + B) \leq \max_\sigma \prod_{i=1}^n (\alpha_i + \beta_{\sigma(i)}), \qquad \text{(VI.54)}$$

where σ varies over all permutations.

Proof. If A and B commute, they can be diagonalised simultaneously, and hence $\det(A + B) = \prod_{i=1}^n (\alpha_i + \beta_{\sigma(i)})$ for some σ. So, the inequality (VI.54) is trivial in this case. Next note that the two extreme sides of (VI.54) are invariant under the transformation $B \to UBU^*$ for every unitary U. Hence, it suffices to prove that for a fixed Hermitian matrix A the function $f(H) = \det(A + H)$ on the unitary orbit $\mathcal{U}_B$ of another Hermitian matrix B attains its minimum and maximum at points that commute with A.

Let B_0 be any extreme point of f on $\mathcal{U}_B$. Then, we must have

$$\left.\frac{d}{dt}\right|_{t=0} \det(A + e^{tK} B_0 e^{-tK}) = 0, \qquad \text{(VI.55)}$$

for every skew-Hermitian K. Now,

$$\det(A + e^{tK} B_0 e^{-tK}) = \det(A + B_0 + t[K, B_0]) + O(t^2).$$

Note that, if X, Y are any two matrices and X is invertible, then

$$\det(X + tY) = \det X(1 + t \operatorname{tr} YX^{-1}) + O(t^2).$$

So, if $A + B_0$ is invertible, the condition (VI.55) reduces to

$$\operatorname{tr}[K, B_0](A + B_0)^{-1} = 0.$$

This is equivalent to saying

$$\operatorname{tr} K(B_0(A+B_0)^{-1} - (A+B_0)^{-1}B_0) = 0.$$

If this is to be true for all skew-Hermitian K, we must have

$$B_0(A+B_0)^{-1} = (A+B_0)^{-1}B_0.$$

Thus B_0 commutes with $(A+B_0)^{-1}$, hence with $A+B_0$, and hence with A.

This proves the theorem under the assumption that $A+B_0$ is invertible. The general case follows from this by a limiting argument. ■

Exercise VI.7.2 *Let A and B be Hermitian matrices. If $\lambda_n^{\downarrow}(A)+\lambda_n^{\downarrow}(B) \geq 0$, then*

$$\prod_{j=1}^{n}(\lambda_j^{\downarrow}(A)+\lambda_j^{\downarrow}(B)) \leq \det(A+B) \leq \prod_{j=1}^{n}(\lambda_j^{\downarrow}(A)+\lambda_j^{\uparrow}(B)). \qquad \text{(VI.56)}$$

This is true, in particular, when A and B are positive matrices.

Theorem VI.7.3 *Let A, B be Hermitian matrices with eigenvalues α_j and β_j, respectively, ordered so that*

$$|\alpha_1| \geq \cdots \geq |\alpha_n| \quad \text{and} \quad |\beta_1| \geq \cdots \geq |\beta_n|.$$

Let $T = A + iB$. Then

$$|\det T| \leq \prod_{j=1}^{n} |\alpha_j + i\beta_{n-j+1}|. \qquad \text{(VI.57)}$$

Proof. The function $f(t) = \frac{1}{2}\log t$ is concave on the positive half-line. Hence, using the majorisation (VI.7) and Corollary II.3.4, we have

$$\sum_{j=1}^{n} \log|\alpha_j + i\beta_{n-j+1}| \geq \sum_{j=1}^{n} \log s_j.$$

Hence,

$$\prod_{j=1}^{n}|\alpha_j + i\beta_{n-j+1}| \geq \prod_{j=1}^{n} s_j = |\det T|.$$

■

Proposition VI.7.4 *Let $T = A + iB$, where A is positive and B Hermitian. Then*

$$|\det T| = \det A \prod_{j=1}^{n}[1 + s_j(A^{-1/2}BA^{-1/2})^2]^{1/2}. \qquad \text{(VI.58)}$$

Proof. Since $T = A^{1/2}(I + iA^{-1/2}BA^{-1/2})A^{1/2}$, we have

$$\det T = \det A \cdot \det(I + iA^{-1/2}BA^{-1/2}). \tag{VI.59}$$

Note that

$$\begin{aligned} & |\det(I + iA^{-1/2}BA^{-1/2})|^2 \\ = \ & \det[(I + iA^{-1/2}BA^{-1/2})(I - iA^{-1/2}BA^{-1/2})] \\ = \ & \det[I + (A^{-1/2}BA^{-1/2})^2] \\ = \ & \prod_{j=1}^{n} [1 + s_j(A^{-1/2}BA^{-1/2})^2]. \end{aligned} \tag{VI.60}$$

So, (VI.58) follows from (VI.59) and (VI.60). ■

Corollary VI.7.5 *If the matrix A in the Cartesian decomposition $T = A + iB$ is positive, then $|\det T| \geq \det A$.*

Theorem VI.7.6 *Let $T = A + iB$, where A and B are positive matrices with eigenvalues $\alpha_1 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \cdots \geq \beta_n$, respectively. Then,*

$$|\det T| \geq \prod_{j=1}^{n} |\alpha_j + i\beta_j|. \tag{VI.61}$$

Proof. We may assume, without loss of generality, that both A, B are positive definite. Because of relations (VI.59) and (VI.60), the theorem will be proved if we show

$$\prod_{j=1}^{n} [1 + s_j(A^{-1/2}BA^{-1/2})^2] \geq \prod_{j=1}^{n} (1 + \alpha_j^{-2}\beta_j^2).$$

Note that

$$s_j(A^{-1/2}BA^{-1/2}) = s_j(A^{-1/2}B^{1/2})^2.$$

From (III.20) we have

$$\{\log s_{n-j+1}(A^{-1/2}) + \log s_j(B^{1/2})\}_j \prec \{\log s_j(A^{-1/2}B^{1/2})\}_j.$$

This is the same as saying

$$\{\log(\alpha_j^{-1/2}\beta_j^{1/2})\}_j \prec \{\log s_j(A^{-1/2}B^{1/2})\}_j.$$

Since the function $\log(1 + e^{4t})$ is convex in t, using Corollary II.3.4 we obtain from the last majorisation

$$\begin{aligned} \sum_{j=1}^{n} \log(1 + \alpha_j^{-2}\beta_j^2) \ & \leq \ \sum_{j=1}^{n} \log\left(1 + s_j(A^{-1/2}B^{1/2})^4\right) \\ & = \ \sum_{j=1}^{n} \log(1 + s_j(A^{-1/2}BA^{-1/2})^2). \end{aligned}$$

This gives the desired inequality. ■

Exercise VI.7.7 *Show that if only one of the two matrices A and B is positive, then the inequality (VI.61) is not necessarily true.*

A very natural generalisation of Theorem VI.7.1 would be the following statement. If A and B are normal then $\det(A+B)$ lies in the convex hull of the products $\prod_{i=1}^{n}(\alpha_i + \beta_{\sigma(i)})$. This is called the **Marcus-de Oliviera Conjecture** and is a well-known open problem in matrix theory.

VI.8 Problems

Problem VI.8.1. Let A be a Hermitian and B a skew-Hermitian matrix. Show that

$$\|\mathrm{Eig}^{|\downarrow|}(A) - \mathrm{Eig}^{|\uparrow|}(B)\|_Q \leq \|A - B\|_Q$$

for every Q-norm.

Problem VI.8.2. Let $T = A + iB$, where A and B are Hermitian. Show that, for $2 \leq p \leq \infty$,

$$\|T\|_p \leq 2^{1-2/p} \min_{\sigma} \|\mathrm{Eig}\, A + \mathrm{Eig}_\sigma(iB)\|_p,$$

$$\max_{\sigma} \|\mathrm{Eig}\, A + \mathrm{Eig}_\sigma(iB)\| \leq 2^{1/2-1/p}\|T\|_p,$$

and that for $1 \leq p \leq 2$,

$$\|T\|_p \leq 2^{1/p-1/2} \min_{\sigma} \|\mathrm{Eig}\, A + \mathrm{Eig}_\sigma(iB)\|_p,$$

$$\max_{\sigma} \|\mathrm{Eig}\, A + \mathrm{Eig}_\sigma(iB)\|_p \leq 2^{2/p-1}\|T\|_p.$$

Problem VI.8.3. A different proof of Theorem VI.3.11 is outlined below. Fill in the details.

Let $n \geq 3$ and let A, B be $n \times n$ unitary matrices. Assume that the eigenvalues α_j and β_j are distinct and the distances $|\alpha_i - \beta_j|$ are also distinct. If γ_1, γ_2 are two points on the unit circle, we write $\gamma_1 < \gamma_2$ if the minor arc from γ_1 to γ_2 goes counterclockwise. We write $(\alpha\beta\gamma)$ if the points α, β, γ on the unit circle are in counterclockwise cyclic order. Number the indices modulo n, e.g., $\alpha_{n+1} = \alpha_1$.

Label the eigenvalues of A so as to have the order $(\alpha_1\alpha_2 \cdots \alpha_n)$. Let $\delta = d(\sigma(A), \sigma(B))$. Assume that $\delta < 2$; otherwise, there is nothing to prove. Label the eigenvalues of B as $\beta_1, \ldots, \beta_n$ in such a way that for any subset J of $\{1, 2, \ldots, n\}$ and for any permutation σ

$$\max_{i \in J} |\alpha_i - \beta_i| \leq \max_{i \in J} |\alpha_i - \beta_{\sigma(i)}|.$$

Then $\delta = \max_{1\le i\le n} |\alpha_i - \beta_i|$. Assume, without loss of generality, that this maximum is attained at $i = 1$ and that $\alpha_1 < \beta_1$. Check the following.

(i) If $\beta_i < \alpha_i$, then neither $(\alpha_1\beta_i\beta_1)$ nor $(\alpha_1\alpha_i\beta_1)$ is possible.

(ii) There exists j such that $|\alpha_{j+1} - \beta_j| > \delta$. Choose and fix one such j.

(ii) We have $(\alpha_1\beta_1\beta_j\alpha_{j+1})$.

(iv) For $1 < i < j$ we have $(\beta_1\beta_i\beta_j)$.

Let K_A be the arc from α_{j+1} positively to α_1 and K_B the arc from β_1 positively to β_j. Then there are $n - j + 1$ of the α_i in K_A and j of the β_i in K_B. Use Proposition VI.3.5 now.

Problem VI.8.4. Let $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ be any complex numbers. Show that there is a number γ such that

$$\max_i|\alpha_i - \gamma| + \max_j|\beta_j - \gamma| \le \sqrt{2}\max_{i,j}|\alpha_i - \beta_j|.$$

(The proof might be long but is not too difficult.) Use this to get another proof of Theorem VI.3.14.

Problem VI.8.5. Let A be a Hermitian and B a normal matrix. If the eigenvalues α_j of A are enumerated as $\alpha_1 \ge \cdots \ge \alpha_n$ and if the eigenvalues β_j of B are enumerated so that Re $\beta_1 \ge \cdots \ge$ Re β_n then

$$\max_{1\le j\le n} |\alpha_j - \beta_j| \le \sqrt{2}\, \|A - B\|.$$

Problem VI.8.6. Let A be a normal matrix with eigenvalues $\alpha_1, \ldots, \alpha_n$. Let B be any other matrix and let $\varepsilon = \|A - B\|$. By Theorem VI.3.3, all the eigenvalues of B are contained in the set $D = \bigcup_j \overline{D}(\alpha_j, \varepsilon)$. Use the argument in the proof of Theorem VI.5.1 to show that each connected component of D contains as many eigenvalues of B as of A. Use this and the Matching Theorem (Theorem II.2.1) to show that

$$d(\sigma(A),\ \sigma(B)) \le (2n - 1)\|A - B\|.$$

[If A and B are both normal, this argument together with the result of Problem II.5.10 shows that $d(\sigma(A), \sigma(B)) \le n\|A - B\|$. However, in this case, the Hoffman-Wielandt inequality gives a stronger result: $d(\sigma(A),\ \sigma(B)) \le \sqrt{n}\, \|A - B\|$. We will see in the next chapter that, in fact, $d(\sigma(A), \sigma(B)) \le 3\|A - B\|$ in this case.]

Problem VI.8.7. Let A be a Hermitian matrix with eigenvalues $\alpha_1 \ge \cdots \ge \alpha_n$, and let B be any matrix with eigenvalues β_j arranged so that

$\text{Re}\beta_1 \geq \cdots \geq \text{Re}\beta_n$. Let $\text{Re}\beta_j = \mu_j$ and $\text{Im}\beta_j = \nu_j$. Choose an orthonormal basis in which B is upper triangular and

$$B = M + iN + iR,$$

where $M = \text{diag}(\mu_1, \ldots, \mu_n)$, $N = \text{diag}(\nu_1, \ldots, \nu_n)$ and R is strictly upper triangular. Show that

$$\|\text{Im}(A - B)\|_2^2 = \|N\|_2^2 + 1/2\|R\|_2^2.$$

Hence,

$$\Sigma|\nu_j|^2 \leq \|\text{Im}(A - B)\|_2^2.$$

Show that

$$(\Sigma(\alpha_j - \mu_j)^2)^{1/2} \leq \|\text{Re}(A - B)\|_2 + \frac{1}{\sqrt{2}}\|R\|_2.$$

Combine the inequalities above to obtain

$$(\Sigma|\alpha_j - \beta_j|^2)^{1/2} \leq \sqrt{2}\, \|A - B\|_2.$$

Compare this with the result of Problem VI.8.5; note that there B was assumed to be normal.

Problem VI.8.8. It follows from the result of the above problem that if A is Hermitian and B an arbitrary matrix, then

$$d(\sigma(A),\, \sigma(B)) \leq \sqrt{2n}\, \|A - B\|.$$

The factor $\sqrt{2n}$ here can be replaced by another that grows only like $\log n$. For this one needs the following fact, which we state without proof. (See the discussion in the Notes at the end of the chapter.)

Let Z be an $n \times n$ matrix whose eigenvalues are all real. Then

$$\|Z - Z^*\| \leq \gamma_n \|Z + Z^*\|,$$

where

$$\gamma_n = \frac{2}{n} \sum_{j=1}^{[n/2]} \cot \frac{2j - 1}{2n}\, \pi.$$

The constant γ_n is the smallest one for which the above norm inequality is true. Approximating the sum by integrals, it is easy to see that $\gamma_n / \log n$ approaches $2/\pi$ as $n \to \infty$.

Using the notations of Problem 7, show that

$$\max|\nu_j| \leq \|\text{Im}\, B\| = \|\text{Im}\,(A - B)\|,$$

$$\max|\alpha_j - \mu_j| \leq \|A - M\| = \|\text{Re}(A - B)\| + \frac{1}{2}\|R - R^*\|.$$

Let $Z = N + R$. Then Z has only real eigenvalues, $Z - Z^* = R - R^*$, and $Z + Z^* = 2\,\mathrm{Im}(A - B)$. Hence,

$$\max |\alpha_j - \beta_j| \leq \|\mathrm{Re}(A - B)\| + (\gamma_n + 1)\|\mathrm{Im}(A - B)\|.$$

This shows that

$$d(\sigma(A),\, \sigma(B)) \leq (\gamma_n + 2)\|A - B\|.$$

Problem VI.8.9. Let A be the Hermitian matrix with entries

$$\begin{array}{lllll} a_{ij} & = & \frac{1}{|i-j|} & \text{if} & i \neq j, \\ a_{ii} & = & 0 & \text{for all} & i. \end{array}$$

Let $B = A + C$ where C is the skew-Hermitian matrix with entries

$$\begin{array}{lllll} c_{ij} & = & \frac{1}{i-j} & \text{if} & i \neq j \\ c_{ii} & = & 0 & \text{for all} & i. \end{array}$$

Then B is strictly lower triangular, hence all its eigenvalues are 0.

Show that $\|A - B\| \leq \pi$ for all n, and $\|A\| = O(\log\, n)$. (This needs some work.) Since A is Hermitian, this means that its spectral radius is $O(\log\, n)$. Thus, in this example, $d(\sigma(A),\, \sigma(B)) = O(\log\ n)$ and $\|A - B\| \leq \pi$. So, the bound obtained in Problem 8 is not too loose.

Problem VI.8.10. For any matrix A, let A_D denote its diagonal part and A_L, A_U its parts below and above the diagonal. Thus $A = A_L + A_D + A_U$. Show that if A is an $n \times n$ normal matrix, then

$$\|A_L\|_2 \leq \sqrt{n-1}\|A_U\|_2, \quad \|A_U\|_2 \leq \sqrt{n-1}\; \|A_L\|_2.$$

The example in VI.6.11 shows that this inequality is sharp.

Problem VI.8.11. Let A be a normal and B an arbitrary $n \times n$ matrix. Choose an orthonormal basis in which B is upper triangular. In this basis write $A = A_L + A_D + A_U$, $\quad B = B_D + B_U$. By the Hoffman-Wielandt Theorem

$$d_2(\sigma(A), \sigma(B)) \leq \|A - B_D\|_2 = \|A - B + B_U\|_2.$$

Note that

$$A - B + B_U = (A - B)_L + (A - B)_D + A_U.$$

Use the result of Problem VI.8.10 to show that

$$\|A - B + B_U\|_2 \leq \sqrt{n}\; \|A - B\|_2.$$

Hence, we have, for A normal and B arbitrary,

$$d_2(\sigma(A), \sigma(B)) \leq \sqrt{n}\; \|A - B\|_2.$$

From this, we get for the Schatten p-norms

$$\begin{aligned} d_p(\sigma(A),\sigma(B)) &\leq n^{1/p}\,\|A-B\|_p, \quad 1\leq p\leq 2 \\ d_p(\sigma(A),\sigma(B)) &\leq n^{1-1/p}\,\|A-B\|_p, \quad 2\leq p\leq \infty. \end{aligned}$$

Show that for $1\leq p\leq 2$ these inequalities are sharp. (See Example VI.6.11.) For $p=\infty$, this gives

$$d(\sigma(A),\ \sigma(B))\leq n\|A-B\|,$$

which is an improvement on the result of Problem 6 above.

If A is Hermitian, then $\|A_U\|_2=\|A_L\|_2$. Using this one obtains a slightly different proof of the last inequality in Problem VI.8.7.

Problem VI.8.12. Let A be an $n\times n$ Hermitian matrix partitioned as $A=\begin{pmatrix} M & R^* \\ R & N\end{pmatrix}$, where M is a $k\times k$ matrix. Let the eigenvalues of A be $\lambda_1\geq\cdots\geq\lambda_n$, those of M be $\mu_1\geq\cdots\geq\mu_k$, and let the singular values of R be $\rho_1\geq\rho_2\geq\cdots$. Show that there exist indices $1\leq i_1<\cdots<i_k\leq n$ such that for every symmetric gauge function Φ we have

$$\Phi(\mu_1-\lambda_{i_1},\ldots,\mu_k-\lambda_{i_k})\leq\Phi(\rho_1,\rho_1,\rho_2,\rho_2,\ldots).$$

In other words, for every unitarily invariant norm we have

$$|||\mathrm{diag}(\mu_1-\lambda_{i_1},\ldots,\mu_k-\lambda_{i_k})|||\leq|||R\oplus R|||.$$

In particular, we have

$$\|\mathrm{diag}(\mu_1-\lambda_{i_1},\ldots,\mu_k-\lambda_{i_k})\|\leq\|R\|$$

and

$$\|\mathrm{diag}(\mu_1-\lambda_{i_1},\ldots,\mu_k-\lambda_{i_k})\|_2\leq\sqrt{2}\,\|R\|_2.$$

Use an argument similar to the one in the proof of Theorem VI.4.1 to show that the factor $\sqrt{2}$ in the last inequality can be replaced by 1. This raises the question whether we have

$$|||\mathrm{diag}(\mu_1-\lambda_{i_1},\ldots,\mu_k-\lambda_{i_k})|||\leq|||R|||$$

for all unitarily invariant norms. This is not so, as can be seen from the example

$$A=\begin{pmatrix} 0 & 1 & \sqrt{3} \\ 1 & 0 & 0 \\ \sqrt{3} & 0 & 0\end{pmatrix}.$$

Problem VI.8.13. Let Φ be a closed subset of $\mathbb{C}$, and let F be a retraction onto Φ. Let $\mathbf{N}(\Phi)$ be the set of all normal matrices whose spectrum is

contained in Φ. Show that if Φ is a convex set, then for every unitarily invariant norm

$$|||B - F(B)||| \leq |||B - A|||,$$

whenever B is a normal matrix and $A \in \mathbf{N}(\Phi)$. If Φ is any closed set, then the above inequality is true for all Q-norms.

Problem VI.8.14. The aim of this problem and the next, is to outline an alternative approach to the normal path inequality (VI.44). This uses slightly more sophisticated notions of differential geometry.

Let A be any $n \times n$ matrix, and let O_A be the orbit of A under the action of the group $\mathbf{GL(n)}$, i.e.,

$$O_A = \{gAg^{-1} : g \in \mathbf{GL(n)}\}.$$

This is called the **similarity orbit** of A; it is the set of all matrices similar to A. Every differentiable curve in O_A passing through A can be parametrised locally as $e^{tX}Ae^{-tX}$, $X \in \mathbf{M(n)}$. By the same argument as in Section VI.4, the tangent space to O_A at the point A can be characterised as

$$T_A O_A = \{[A, X] : X \in \mathbf{M(n)}\}.$$

The orthogonal complement of this space in $\mathbf{M(n)}$ can be calculated as in Lemma VI.4.2. Show that

$$(T_A O_A)^{\perp} = Z(A^*).$$

Now, a matrix A is normal if and only if $Z(A^*) = Z(A)$. So, for a normal matrix we have a direct sum decomposition

$$\mathbf{M(n)} = T_A O_A \oplus Z(A).$$

Now, if $B \in O_A$, then B and A have the same set of eigenvalues and hence

$$d_2(\sigma(A),\ \sigma(B)) = 0.$$

If $B \in Z(A)$, then there is an orthonormal basis in which A and B are both upper triangular. Hence, for such a B,

$$d_2(\sigma(A),\ \sigma(B)) \leq \|A - B\|_2.$$

Now, let $\gamma(t), 0 \leq t \leq 1$ be a C^1 curve in the space of normal matrices. Let $\gamma(0) = A_0, \gamma(1) = A_1$. Let $\varphi(A) = d_2(\sigma(A_0), \sigma(A))$. At each point $\gamma(t)$ consider the decomposition

$$\mathbf{M(n)} = T_{\gamma(t)} O_{\gamma(t)} \oplus Z(\gamma(t))$$

obtained above. Then, as we move along $\gamma(t)$, the rate of change of the function φ is zero in the first direction in this decomposition, and in the

second, it is bounded by the rate of change of the argument. Hence we should have

$$\varphi(\gamma(1)) \leq \int_0^1 \|\gamma'(t)\|_2 dt.$$

Prove this. Note that this says that

$$d_2(\sigma(A_0),\ \sigma(A_1)) \leq \ell_2(\gamma),$$

if γ is a C^1 curve passing through normal matrices and joining A_0 to A_1.

Problem VI.8.15. The two crucial properties of the Frobenius norm used above were its invariance under the conjugations $A \to UAU^*$ and the pinching inequality. The first made it possible to change to any orthonormal basis, and the second was used to conclude that the diagonal of a matrix has norm smaller than the whole matrix. Both these properties are enjoyed by all wui norms. So, the method outlined above can be adopted to work for all wui norms to give the same result. (Some conditions on the path are necessary to ensure differentiability of the functions involved.)

Problem VI.8.16. Fill in the details in the following outline of a proof of the statement: every complex matrix with trace 0 is a commutator of two matrices.

Let A be a matrix such that $\text{tr}A = 0$. Assume that A is upper triangular. Let B be the nilpotent upper Jordan matrix (i.e., B has all entries 0 except the ones on the first superdiagonal, which are all 1). Then $Z(B^*)$ contains only polynomials in B^*. (This is a general fact: $Z(X)$ contains only polynomials in X if and only if in the Jordan form of X there is just one block for each different eigenvalue.) Thus $Z(B^*)$ consists of lower triangular matrices with constant diagonals. Show that A is orthogonal to all such matrices. Hence A is in the space $T_B O_B$, and so $A = [B, C]$ for some C.

VI.9 Notes and References

Perturbation theory for eigenvalues is of interest to mathematicians, physicists, engineers, and numerical analysts. Among the several books that deal with this topic are the venerable classics, T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, 1966, and J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965. The first is addressed to the problems of quantum physics, the second to those of numerical analysis. *Matrix Computations* by G.H. Golub and C.F. Van Loan, The Johns Hopkins University Press, 1983, has enough of interest for the theorist and for the designer of algorithms. Much closer to the spirit of our book (but a lot more appealing to the numerical analyst) is *Matrix Perturbation Theory*

by G.W. Stewart and J.-G. Sun, Academic Press, 1990. Much of the material in this chapter has appeared before in R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman, 1987.

In Section VI.1, we have done the bare minimum of qualitative analysis required for the latter sections. The questions of differentiability and analyticity of eigenvalues and eigenvectors, asymptotic expansions and their convergence, and other related matters are discussed at great length by T. Kato, and by H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser, 1984. See also M. Reed and B. Simon, *Analysis of Operators*, Academic Press, 1978.

The proof of Theorem VI.1.2 given here is adopted from H. Whitney, *Complex Analytic Varieties*, Addison Wesley, 1972. Other proofs may be found in R. Bhatia and K.K. Mukherjea, *The space of unordered tuples of complex numbers*, Linear Algebra Appl., 52/53 (1983) 765-768. The proof of Theorem VI.1.4 is taken from T. Kato.

Theorem VI.4.1 was proved in A.J. Hoffman and H.W. Wielandt, *The variation of the spectrum of a normal matrix*, Duke Math J., 20 (1953) 37-39. This was believed to be the next best thing to having an analogue of Weyl's Perturbation Theorem for normal matrices . The conjecture, that for normal A, B the optimal matching distance in each unitarily invariant norm is bounded by the corresponding distance between A and B, seems to have been raised first in L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quarterly J. Math. 11 (1960) 50-59. The problem, of course, turned out to be much more subtle than imagined at that time.

Theorem VI.2.2 is due to V.S. Sunder, *Distance between normal operators*, Proc. Amer. Math. Soc., 84 (1982) 483-484. The rest of Section 2 is based on T. Ando and R. Bhatia, *Eigenvalue inequalities associated with the Cartesian decomposition*, Linear and Multilinear Algebra, 22 (1987) 133-147.

Theorem VI.3.3 is one of the several results proved in the paper *Norms and exclusion theorems*, Numer. Math. 2(1960) 137-141, by F.L. Bauer and C.T. Fike. Theorem VI.3.11 was first proved by R. Bhatia and C. Davis, *A bound for the spectral variation of a unitary operator*, Linear and Multilinear Algebra, 15 (1984) 71-76. Their proof is summarised in Problem VI.8.3. This approach of ordering eigenvalues in the cyclic order is elaborated further by L. Elsner and C. He, *Perturbation and interlace theorems for the unitary eigenvalue problem*, Linear Algebra Appl., 188/189 (1993) 207-229, where many related results are proved. The proof given in Section 3 is adapted from R.H. Herman and A. Ocneanu, *Spectral analysis for automorphisms of UHF C^*-algebras*, J. Funct. Anal. 66 (1986) 1-10.

The first example of 3×3 normal matrices A, B for which $d(\sigma(A), \sigma(B)) > \|A - B\|$ was constructed by J.A. R. Holbrook, *Spectral variation of normal matrices*, Linear Algebra Appl., 174 (1992) 131-144. This was done using a directed computer search. The bare-hands example VI.3.13 was shown to us by G.M. Krause. The inequality (VI. 33) appears in T. Ando, *Bounds*

for the antidistance, J. Convex Analysis, 2(1996) 1-3. It had been observed earlier by R. Bhatia that this would lead to Theorem VI.3.14. A different proof of this theorem was found independently by M. Omladic and P. Semrl, *On the distance between normal matrices*, Proc. Amer. Math. Soc., 110 (1990) 591-596. The idea of the simpler proof in Problem VI.8.4 is due to L. Elsner.

The geometric ideas of Sections VI.5 and VI.6 were introduced in R. Bhatia, *Analysis of spectral variation and some inequalities*, Trans. Amer. Math. Soc., 272 (1982) 323-332. They were developed further in three papers by R. Bhatia and J.A.R. Holbrook: *Short normal paths and spectral variation*, Proc. Amer. Math. Soc., 94 (1985) 377-382; *Unitary invariance and spectral variation*, Linear Algebra Appl., 95 (1987) 43-68; and *A softer, stronger Lidskii theorem*, Proc. Indian Acad. Sci. (Math. Sci.), 99(1989) 75-83. Much of Sections 5 and 6 is based on these papers.

Example VI.5.8 is due to M.D. Choi. The special operator norm case of the inequality (VI.49) was proved by K.R. Parthasarathy, *Eigenvalues of matrix-valued analytic maps*, J. Austral. Math. Soc. (Ser. A), 26(1978) 179-197. Theorem VI.6.10 and Example VI.6.11 are taken from R. Bhatia, C. Davis and A. McIntosh, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52/53 (1983) 45-67. The inequality (VI.53) for (strongly) unitarily invariant norms was first proved by V.S. Sunder, *On permutations, convex hulls and normal operators*, Linear Algebra Appl., 48 (1982) 403-411.

P.R. Halmos, *Spectral approximants of normal operators*, Proc. Edinburgh Math. Soc. 19 (1974) 51-58, initiated the study of operator approximation problems of the following kind. Given a normal operator A, find the operator closest to A from the class of normal operators that have their spectrum in a given closed set Φ. The result of Theorem VI.6.2 for infinite-dimensional Hilbert spaces (and for closed sets Φ) was proved in this paper for the special case of the operator norm. This was extended to Schatten p-norms, $p \geq 2$, by R. Bouldin, *Best approximation of a normal operator in the Schatten p-norm*, Proc. Amer. Math. Soc., 80 (1980) 277-282. The result for Q-norms, as well as the special one for all unitarily invariant norms given in Problem VI.8.11, was proved by R. Bhatia, *Some inequalities for norm ideals*, Commun. Math. Phys. 111(1987) 33-39.

Theorem VI.7.1 is due to M. Fiedler, *Bounds for the determinant of the sum of Hermitian matrices*, Proc. Amer. Math. Soc., 30(1971) 27-31. The inequality (VI.56) is also proved in this paper. Theorem VI.7.3 is due to J.F. Queiró and A.L. Duarte, *On the Cartesian decomposition of a matrix*, Linear and Multilinear Algebra, 18 (1985) 77-85, while Theorem VI.7.6 is due to N. Bebiano. The proofs given here are taken from the paper by T. Ando and R. Bhatia cited above. There are several papers related to the Marcus-de Oliviera conjecture. For a recent survey, see N. Bebiano, *New developments on the Marcus-Oliviera Conjecture*, Linear Algebra Appl., 197/198 (1994) 793-802.

The results in Problem VI.8.7 are due to W. Kahan, *Spectra of nearly Hermitian matrices*, Proc. Amer. Math. Soc., 48 (1975) 11-17, as are the ones in Problem VI.8.8 (essentially) and Problem VI.8.9. In an earlier paper, *Every $n \times n$ matrix Z with real spectrum satisfies* $\|Z - Z^*\| \leq \|Z + Z^*\|(\log_2 n + 0.038)$, Proc. Amer. Math. Soc. 39(1973) 235-241, Kahan proved the inequality in the title of the paper and used it to obtain the perturbation bound in the later paper. The exact value of γ_n in this inequality was obtained by A. Pokrzywa, *Spectra of operators with fixed imaginary parts*, Proc. Amer. Math. Soc., 81(1981) 359-364. The appearance of a constant growing like $\log n$ in this inequality is related to another important problem. Let $\mathcal{T}$ be the linear operator on $\mathbf{M(n)}$ that takes every matrix to its upper triangular part. Then sup $\|\mathcal{T}(X)\|/\|X\|$ is known to be $O(\log\ n)$. From this one can see (on reducing Z to an upper triangular form) that the constant γ_n also must have this order. Related results may be found in R. Mathias, *The Hadamard operator norm of a circulant and applications*, SIAM J. Matrix Anal. Appl., 14(1993) 1152-1167.

The results in Problems VI.8.10 and VI.8.11 are taken from J.-G Sun, *On the variation of the spectrum of a normal matrix*, Linear Algebra Appl., 246(1996) 215-223.

Bounds like the one given in Problem VI.8.12 are called *residual bounds* in numerical analysis. See W. Kahan, *Numerical linear algebra*, Canadian Math. Bull., 9 (1966) 757-801, where the special improvement for the Frobenius norm is obtained. The general result for all unitarily invariant norms is proved in the book by Stewart and Sun. The example at the end of this problem was given in R. Bhatia, *On residual bounds for eigenvalues*, Indian J. Pure Appl. Math., 23 (1992) 865-866. The result in Problem VI.8.14 is a well-known theorem; the proof we have outlined here was shown to us by V.S. Sunder.

VII

Perturbation of Spectral Subspaces of Normal Matrices

In Chapter 6 we saw that the eigenvalues of a (normal) matrix change continuously with the matrix. The behaviour of eigenvectors is more complicated. The following simple example is instructive. Let $A = \begin{pmatrix} 1+\varepsilon & 0 \\ 0 & 1-\varepsilon \end{pmatrix} \oplus H$, $B = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix} \oplus H$, where H is Hermitian. The eigenvalues of the first 2×2 block of A are $1+\varepsilon, 1-\varepsilon$. The same is true for B. The corresponding normalised eigenvectors are $(1,0)$ and $(0,1)$ for A, and $\frac{1}{\sqrt{2}}(1,1)$ and $\frac{1}{\sqrt{2}}(1,-1)$ for B. As $\varepsilon \to 0, B$ and A approach each other, but their eigenvectors remain stubbornly apart. Note, however, that the *eigenspaces* that these two eigenvectors of A and B span are identical. In this chapter we will see that interesting and useful perturbation bounds may be obtained for eigenspaces corresponding to closely bunched eigenvalues of normal matrices.

Before we do this, it is necessary to introduce notions of distance between two subspaces. Also, it turns out that this perturbation problem is closely related to the solution of the matrix equation $AX - XB = Y$. This equation called the **Sylvester Equation**, arises in several other contexts. So, we will study it in some detail before applying the results to the perturbation problem at hand.

VII.1 Pairs of Subspaces

We will be dealing with block decompositions of matrices. To keep track of dimensions, we will find it convenient to write

$$A = \begin{array}{c} \begin{matrix} k & \ell \end{matrix} \\ \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \end{array} \begin{matrix} m \\ p \end{matrix}$$

for a block-matrix in which k and ℓ are the number of columns, and m and p are the number of rows in the blocks indicated.

Theorem VII.1.1 *(The QR Decomposition) Let A be an $m \times n$ matrix, $m \geq n$. Then there is an $m \times m$ unitary matrix Q such that*

$$Q^*A = \begin{array}{c} n \\ \begin{pmatrix} R \\ 0 \end{pmatrix} \end{array} \begin{matrix} n \\ m-n \end{matrix}, \tag{VII.1}$$

where R is upper triangular with nonnegative real diagonal entries.

Proof. For a square matrix A, this was proved in Chapter 1. The same proof also works here. (In essence this is just the Gram-Schmidt process.) ■

The matrix R above is called the **R factor** of A.

Exercise VII.1.2 *Let A be an $m \times n$ matrix with rank $A = n$. Then the R factor in the QR decomposition of A has positive diagonal elements and is uniquely determined. (See Exercise I.2.2.) If we write*

$$Q = \begin{array}{c} \begin{matrix} n & m-n \end{matrix} \\ \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \end{array} m,$$

then we have

$$A = Q_1 R, \quad Q_1 = AR^{-1}.$$

Thus Q_1 is uniquely determined by A. However, Q_2 need not be unique. Note the range of A is the range of Q_1, and its orthogonal complement is the range of Q_2.

Exercise VII.1.3 *Let A be an $m \times n$ matrix with $m \leq n$. Then there exists an $n \times n$ unitary matrix W such that*

$$AW = \begin{array}{c} \begin{matrix} m & n-m \end{matrix} \\ \begin{pmatrix} L & 0 \end{pmatrix} \end{array} m,$$

where L is lower triangular and has nonnegative real diagonal entries.

We remark here that it is only for convenience that we choose R (and L) to have nonnegative diagonal entries. By modifying Q (and W), we can also make R (and L) have nonpositive real diagonal entries.

Exercise VII.1.4 *Let A be an $m \times n$ matrix with rank $A = r$. Then there exists an $n \times n$ permutation matrix P and an $m \times m$ unitary matrix Q such that*

$$Q^*AP = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix},$$

where R_{11} is an $r \times r$ upper triangular matrix with positive diagonal entries. This is called a **rank revealing QR decomposition**.

Exercise VII.1.5 *Let A be an $m \times n$ matrix with rank $A = r$. Then there exists an $m \times m$ unitary matrix Q and an $n \times n$ unitary matrix W such that*

$$Q^*AW = \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix},$$

where T is an $r \times r$ triangular matrix with positive diagonal entries.

Theorem VII.1.6 *(The CS Decomposition) Let W be an $n \times n$ unitary matrix partitioned as*

$$W = \begin{matrix} & \ell & m & \\ & \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} & & \begin{matrix} \ell \\ m \end{matrix} \end{matrix} \tag{VII.2}$$

where $\ell \leq m$. Then there exist unitary matrices $U = \operatorname{diag}(U_{11}, U_{22})$ and $V = \operatorname{diag}(V_{11}, V_{22})$, where U_{11}, V_{11} are $\ell \times \ell$ matrices, such that

$$U^*WV = \begin{matrix} \ell \quad\ \ \ell \quad\ m-\ell \\ \begin{pmatrix} C & -S & 0 \\ S & C & 0 \\ 0 & 0 & I \end{pmatrix} \end{matrix} \begin{matrix} \ell \\ \ell \\ m-\ell \end{matrix} \tag{VII.3}$$

where C and S are nonnegative diagonal matrices, with diagonal entries $0 \leq c_1 \leq \cdots \leq c_\ell \leq 1$ and $1 \geq s_1 \geq \cdots \geq s_\ell \geq 0$, respectively, and

$$C^2 + S^2 = I.$$

Proof. For the sake of brevity, let us call a map $X \to U^*XV$ on the space of $n \times n$ matrices a $\mathcal{U}$-transform, if U, V are block diagonal unitary matrices with top left blocks of size $\ell \times \ell$. The product of two $\mathcal{U}$-transforms is again a $\mathcal{U}$-transform. We will prove the theorem by showing that one can change the matrix W in (VII.2) to the form (VII.3) by a succession of $\mathcal{U}$-transforms.

Let U_{11}, V_{11} be $\ell \times \ell$ unitary matrices such that

$$U_{11}^* W_{11} V_{11} = \begin{array}{c} \begin{matrix} k & \ell - k \end{matrix} \\ \begin{pmatrix} C_1 & 0 \\ 0 & I \end{pmatrix} \end{array},$$

where C_1 is a diagonal matrix with diagonal entries $0 \leq c_1 \leq c_2 \leq \cdots \leq c_k < 1$. This is just the singular value decomposition: since W is unitary, all singular values of W_{11} are bounded by 1. Then the $\mathcal{U}$-transform in which $U = \text{diag}\ (U_{11}, I)$ and $V = \text{diag}\ (V_{11}, I)$ reduces W to the form

$$\begin{array}{c|c} & \begin{matrix} k & \ell-k & & m \end{matrix} \\ \begin{matrix} k \\ \ell-k \\ \\ m \end{matrix} & \left(\begin{array}{cc|c} C_1 & 0 & ? \\ 0 & I & ? \\ \hline ? & ? & ? \end{array}\right) \end{array}, \qquad \text{(VII.4)}$$

where the structures of the three blocks whose entries are indicated by ? are yet to be determined. Let W_{21} denote now the bottom left corner of the matrix (VII.4). By the QR Decomposition Theorem we can find an $m \times m$ unitary matrix Q_{22} such that

$$Q_{22}^* W_{21} = \begin{array}{c} \ell \\ \begin{pmatrix} R \\ 0 \end{pmatrix} \end{array} \begin{matrix} \ell \\ m-\ell \end{matrix}, \qquad \text{(VII.5)}$$

where R is upper triangular with nonnegative diagonal entries. The $\mathcal{U}$-transform in which $U = \text{diag}\ (I, Q_{22})$ and $V = \text{diag}\ (I, I)$ leaves the top left corner of (VII.4) unchanged and changes the bottom left corner to the form (VII.5). Assume that this transform has been carried out. Using the fact that the columns of a unitary matrix are orthonormal, one sees that the last $\ell - k$ columns of R must be zero. Now examine the remaining columns, proceeding from left to right. Since C_1 is diagonal with nonnegative diagonal entries, all of which are strictly smaller than 1, one sees that R is also diagonal. Thus the matrix W is now reduced to the form

$$\begin{array}{c|c} & \begin{matrix} k & \ell-k & & m \end{matrix} \\ \begin{matrix} k \\ \ell-k \\ \\ k \\ \ell-k \\ m-\ell \end{matrix} & \left(\begin{array}{cc|c} C_1 & 0 & ? \\ 0 & I & ? \\ \hline S_1 & 0 & ? \\ 0 & 0 & ? \\ 0 & 0 & ? \end{array}\right) \end{array}, \qquad \text{(VII.6)}$$

in which S_1 is diagonal with $C_1^2 + S_1^2 = I$, and hence $0 \leq S_1 \leq I$. The structures of the two blocks on the right are yet to be determined. Now, by

Exercise VII.1.3 and the remark following it, we can find an $m \times m$ unitary matrix V_{22} which on right multiplication converts the top right block of (VII.6) to the form

$$\begin{array}{cc} & \begin{array}{cc} \ell & m-\ell \end{array} \\ \ell & \begin{pmatrix} L & 0 \end{pmatrix}, \end{array} \tag{VII.7}$$

in which L is lower triangular with nonpositive diagonal entries. The $\mathcal{U}$-transform in which $U = \text{diag}\ (I, I)$ and $V = \text{diag}\ (I, V_{22})$ leaves the two left blocks in (VII.6) unchanged and converts the top right block to the form (VII.7). Again, orthonormality of the rows of a unitary matrix and a repetition of the argument in the preceding step show that after this $\mathcal{U}$-transform the matrix W is reduced to the form

$$\begin{array}{c} \\ k \\ \ell-k \\ \\ k \\ \ell-k \\ m-\ell \end{array} \begin{array}{c} \begin{array}{ccccc|ccc} k & \ell-k & & & & k & \ell-k & m-\ell \end{array} \\ \left(\begin{array}{cc|ccc} C_1 & 0 & -S_1 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \hline S_1 & 0 & X_{33} & X_{34} & X_{35} \\ 0 & 0 & X_{43} & X_{44} & X_{45} \\ 0 & 0 & X_{53} & X_{54} & X_{55} \end{array}\right) \end{array}. \tag{VII.8}$$

Now, we determine the form of the bottom right corner. Since the rows of a unitary matrix are mutually orthogonal, we must have $C_1 S_1 = S_1 X_{33}$. But C_1 and S_1 are diagonal and S_1 is invertible. Hence, we must have $X_{33} = C_1$. But then the blocks $X_{34}, X_{35}, X_{43}, X_{53}$ must all be 0, since the matrix (VII.8) is unitary. So, this matrix has the form

$$\begin{array}{c} \\ k \\ \ell-k \\ \\ k \\ \ell-k \\ m-\ell \end{array} \begin{array}{c} \begin{array}{ccccc|ccc} k & \ell-k & & & & k & \ell-k & m-\ell \end{array} \\ \left(\begin{array}{cc|ccc} C_1 & 0 & -S_1 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \hline S_1 & 0 & C_1 & 0 & 0 \\ 0 & 0 & 0 & X_{44} & X_{45} \\ 0 & 0 & 0 & X_{54} & X_{55} \end{array}\right) \end{array}. \tag{VII.9}$$

Let $X = \begin{pmatrix} X_{44} & X_{45} \\ X_{54} & X_{55} \end{pmatrix}$. Then X is a unitary matrix of size $m-k$. Let $U = \text{diag}\ (I_\ell, I_k, X)$, where I_ℓ and I_k are the identity operators of sizes ℓ and k, respectively. Then, multiplying (VII.9) on the left by U^*— another

$\mathcal{U}$-transform — we reduce it to the form

$$\begin{array}{cc} & \begin{array}{ccccccc} k & \ell-k & & k & \ell-k & & m-\ell \end{array} \\ \begin{array}{c} k \\ \ell-k \\ \\ k \\ \ell-k \\ \\ m-\ell \end{array} & \begin{pmatrix} C_1 & 0 & | & -S_1 & 0 & | & 0 \\ 0 & I & | & 0 & 0 & | & 0 \\ -- & -- & -- & -- & -- & | & -- \\ S_1 & 0 & | & C_1 & 0 & | & 0 \\ 0 & 0 & | & 0 & I & | & 0 \\ -- & -- & -- & -- & -- & | & -- \\ 0 & 0 & | & 0 & 0 & | & I \end{pmatrix} \end{array}. \tag{VII.10}$$

If we now put

$$C = \begin{array}{cc} \begin{array}{cc} k & \ell - k \end{array} & \\ \begin{pmatrix} C_1 & 0 \\ 0 & I \end{pmatrix} & \begin{array}{c} k \\ \ell - k \end{array} \end{array}$$

and

$$S = \begin{array}{cc} \begin{array}{cc} k & \ell - k \end{array} & \\ \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} & \begin{array}{c} k \\ \ell - k \end{array}, \end{array}$$

then the matrix (VII.10) is in the desired form (VII.3). ■

Exercise VII.1.7 *Let W be as in (VII.2) but with $\ell \geq m$. Show that there exist unitary matrices $U=$ diag (U_{11}, U_{22}) and $V =$ diag (V_{11}, V_{22}), where U_{11}, V_{11} are $\ell \times \ell$ matrices, such that*

$$U^*WV = \begin{array}{cc} \begin{array}{ccc} n-\ell & 2\ell-n & n-\ell \end{array} & \\ \begin{pmatrix} C & 0 & -S \\ 0 & I & 0 \\ S & 0 & C \end{pmatrix} & \begin{array}{c} n-\ell \\ 2\ell - n \\ n - \ell \end{array}, \end{array} \tag{VII.11}$$

where C and S are nonnegative diagonal matrices with diagonal entries $0 \leq c_1 \leq \cdots \leq c_{n-\ell} \leq 1$ and $1 \geq s_1 \geq \cdots \geq s_{n-\ell} \geq 0$, respectively, and $C^2 + S^2 = I$.

The form of the matrices C and S in the above decompositions suggests an obvious interpretation in terms of angles. There exist (acute) angles θ_j, $\frac{\pi}{2} \geq \theta_1 \geq \theta_2 \geq \cdots \geq 0$, such that $c_j = \cos\theta_j$ and $s_j = \sin\theta_j$.

One of the major applications of the CS decomposition is the facility it provides for analysing the relative position of two subspaces of $\mathbb{C}^n$.

Theorem VII.1.8 *Let X_1, Y_1 be $n \times \ell$ matrices with orthonormal columns. Then there exist $\ell \times \ell$ unitary matrices U_1 and V_1 and an $n \times n$ unitary matrix Q with the following properties.*

(i) If $2\ell \le n$, then

$$QX_1U_1 = \overset{\ell}{\begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix}} \begin{matrix} \ell \\ \ell \\ n-2\ell \end{matrix} , \tag{VII.12}$$

$$QY_1V_1 = \overset{\ell}{\begin{pmatrix} C \\ S \\ 0 \end{pmatrix}} \begin{matrix} \ell \\ \ell \\ n-2\ell \end{matrix} , \tag{VII.13}$$

where C, S are diagonal matrices with diagonal entries $0 \le c_1 \le \cdots \le c_\ell \le 1$ and $1 \ge s_1 \ge \cdots \ge s_\ell \ge 0$, respectively, and $C^2 + S^2 = I$.

(ii) If $2\ell > n$, then

$$QX_1U_1 = \begin{array}{c} \begin{matrix} n-\ell & 2\ell-n \end{matrix} \\ \begin{pmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix} \end{array} \begin{matrix} n-\ell \\ 2\ell-n \\ n-\ell \end{matrix} , \tag{VII.14}$$

$$QY_1V_1 = \begin{array}{c} \begin{matrix} n-\ell & 2\ell-n \end{matrix} \\ \begin{pmatrix} C & 0 \\ 0 & I \\ S & 0 \end{pmatrix} \end{array} \begin{matrix} n-\ell \\ 2\ell-n \\ n-\ell \end{matrix} , \tag{VII.15}$$

where C, S are diagonal matrices with diagonal entries $0 \le c_1 \le \cdots \le c_{n-\ell} \le 1$ and $1 \ge s_1 \ge \cdots \ge s_{n-\ell} \ge 0$, respectively, and $C^2 + S^2 = I$.

Proof. Let $2\ell \le n$. Choose $n \times (n-\ell)$ matrices X_2 and Y_2 such that $X = (X_1 \ \ X_2)$ and $Y = (Y_1 \ \ Y_2)$ are unitary. Let

$$W = X^*Y = \begin{array}{c} \begin{matrix} \ell & \quad n-\ell \end{matrix} \\ \begin{pmatrix} X_1^*Y_1 & X_1^*Y_2 \\ X_2^*Y_1 & X_2^*Y_2 \end{pmatrix} \end{array} \begin{matrix} \ell \\ n-\ell \end{matrix} .$$

By Theorem VII.1.6 we can find block diagonal unitary matrices $U = \mathrm{diag}(U_1, U_2)$ and $V = \mathrm{diag}(V_1, V_2)$, in which U_1 and V_1 are $\ell \times \ell$ unitaries, such that

$$\begin{pmatrix} U_1^*X_1^*Y_1V_1 & U_1^*X_1^*Y_2V_2 \\ U_2^*X_2^*Y_1V_1 & U_2^*X_2^*Y_2V_2 \end{pmatrix} = \begin{pmatrix} C & -S & 0 \\ S & C & 0 \\ 0 & 0 & I \end{pmatrix} .$$

Let $Q = (XU)^* = (X_1U_1 \ \ X_2U_2)^*$. Then from the first columns of the two sides of the above equation we obtain the equation (VII.13). For this Q the equation (VII.12) is also true.

When $2\ell > n$, the assertion of the theorem follows, in the same manner, from the decomposition (VII.11). ■

This theorem can be interpreted as follows. Let $\mathcal{E}$ and $\mathcal{F}$ be ℓ-dimensional subspaces of $\mathbb{C}^n$. Choose orthonormal bases $x_1, \ldots, x_\ell$ and $y_1, \ldots, y_\ell$ for these spaces. Let $X_1 = (x_1\, x_2\, \cdots\, x_\ell)$, $Y_1 = (y_1\, y_2\, \cdots\, y_\ell)$. Premultiplying X_1, Y_1 by Q corresponds to a unitary transformation of the whole space $\mathbb{C}^n$, while postmultiplying X_1 by U_1 and Y_1 by V_1 corresponds to a change of bases within the spaces $\mathcal{E}$ and $\mathcal{F}$, respectively. Thus, the theorem says that, if $2\ell \leq n$, then there exists a unitary transformation Q of $\mathbb{C}^n$ such that the columns of the matrices on the right-hand sides in (VII.12) and (VII.13) form orthonormal bases for $Q\mathcal{E}$ and $Q\mathcal{F}$, respectively. The span of those columns in the second matrix, for which $s_j = 1$, is the orthogonal complement of $Q\mathcal{E}$ in $Q\mathcal{F}$. When $2\ell > n$, the columns of the matrices on the right-hand sides of (VII.14) and (VII.15) form orthonormal bases for $Q\mathcal{E}$ and $Q\mathcal{F}$, respectively. The last $2\ell - n$ columns are orthonormal vectors in the intersection of these two spaces. The space spanned by those columns of the second matrix, in which $s_j = 1$, is the orthogonal complement of $Q\mathcal{E}$ in $Q\mathcal{F}$.

The reader might find it helpful to see what the above theorem says when $\mathcal{E}$ and $\mathcal{F}$ are lines or planes in $\mathbb{R}^3$.

Using the notation above, we set

$$\Theta(\mathcal{E}, \mathcal{F}) = \arcsin S.$$

This is called the **angle operator** between the subspaces $\mathcal{E}$ and $\mathcal{F}$. It is a diagonal matrix, and its diagonal entries are called the **canonical angles** between the subspaces $\mathcal{E}$ and $\mathcal{F}$.

If the columns of a matrix X are orthonormal and span the subspace $\mathcal{E}$, then the orthogonal projection onto $\mathcal{E}$ is given by the matrix $E = XX^*$. This fact is used repeatedly below.

Exercise VII.1.9 *Let $\mathcal{E}$ and $\mathcal{F}$ be subspaces of $\mathbb{C}^n$. Let X and Y be matrices with orthonormal columns that span $\mathcal{E}$ and $\mathcal{F}$, respectively. Let E, F be the orthogonal projections with ranges $\mathcal{E}, \mathcal{F}$. Then the nonzero singular values of EF are the same as the nonzero singular values of X^*Y.*

Exercise VII.1.10 *Let $\mathcal{E}, \mathcal{F}$ be subspaces of $\mathbb{C}^n$ of the same dimension, and let E, F be the orthogonal projections with ranges $\mathcal{E}, \mathcal{F}$. Then the singular values of EF are the cosines of the canonical angles between $\mathcal{E}$ and $\mathcal{F}$, and the nonzero singular values of $E^\perp F$ are the sines of the nonzero canonical angles between $\mathcal{E}$ and $\mathcal{F}$.*

Exercise VII.1.11 *Let $\mathcal{E}, \mathcal{F}$ and E, F be as above. Then the nonzero singular values of $E - F$ are the nonzero singular values of $E^\perp F$, each counted twice; i.e., these are the numbers $s_1, s_1, s_2, s_2, \ldots$.*

Note that by Exercise VII.1.10, the angle operator $\Theta(\mathcal{E},\mathcal{F})$ does not depend on the choice of any particular bases in $\mathcal{E}$ and $\mathcal{F}$. Further, $\Theta(\mathcal{E},\mathcal{F}) = \Theta(\mathcal{F},\mathcal{E})$.

It is natural to define the **distance between the spaces** $\mathcal{E}$ and $\mathcal{F}$ as $\|E-F\|$. In view of Exercise VII.1.11, this is also the number $\|E^{\perp}F\|$. More generally, we might consider $|||E^{\perp}F|||$, for every unitarily invariant norm, to define a distance between the spaces $\mathcal{E}$ and $\mathcal{F}$. In this case, $|||E-F||| = |||E^{\perp}F \oplus EF^{\perp}|||$.

We could use the numbers $|||E^{\perp}F|||$ to measure the separation of $\mathcal{E}$ and $\mathcal{F}$, even when they have different dimensions. Even the principal angles can be defined in this case:

Exercise VII.1.12 *Let x, y be any two vectors in $\mathbb{C}^n$. The angle between x and y is defined to be a number $\angle(x,y)$ in $[0,\pi/2]$ such that*

$$\angle(x,y) = \cos^{-1}\frac{|y^*x|}{\|x\|\;\|y\|}.$$

Let $\mathcal{E}$ and $\mathcal{F}$ be subspaces of $\mathbb{C}^n$, and let $\dim\mathcal{E} \geq \dim\mathcal{F} = m$. *Define $\theta_1,\ldots,\theta_m$ recursively as*

$$\theta_k = \max_{\substack{x\in\mathcal{E}\\ x\perp[x_1,\ldots,x_{k-1}]}} \min_{\substack{y\in\mathcal{F}\\ y\perp[y_1,\ldots,y_{k-1}]}} \angle(x,y) = \angle(x_k,y_k).$$

Then $\frac{\pi}{2} \geq \theta_1 \geq \cdots \geq \theta_m \geq 0$. The numbers θ_k are called the principal angles between $\mathcal{E}$ and $\mathcal{F}$. Show that when $\dim\mathcal{E} = \dim\mathcal{F}$, *this coincides with the earlier definition of principal angles.*

Exercise VII.1.13 *Show that for any two orthogonal projections E, F we have $\|E-F\| \leq 1$.*

Proposition VII.1.14 *Let E, F be two orthogonal projections such that $\|E-F\| < 1$. Then the ranges of E and F have the same dimensions.*

Proof. Let $\mathcal{E},\mathcal{F}$ be the ranges of E and F. Suppose $\dim\mathcal{E} > \dim\mathcal{F}$. We will show that $\mathcal{E}\cap\mathcal{F}^{\perp}$ contains a nonzero vector. This will show that $\|E-F\| = 1$.

Let $\mathcal{G} = E\mathcal{F}$. Then $\mathcal{G}\subset\mathcal{E}$, and $\dim\mathcal{G} \leq \dim\mathcal{F} < \dim\mathcal{E}$. Hence $\mathcal{E}\cap\mathcal{G}^{\perp}$ contains a nonzero vector x. It is easy to see that $\mathcal{E}\cap\mathcal{G}^{\perp}\subset\mathcal{F}^{\perp}$. Hence, $x\in\mathcal{F}^{\perp}$. ■

In most situations in perturbation theory we will be interested in comparing two projections E, F such that $\|E-F\|$ is small. The above proposition shows that in this case $\dim E = \dim F$.

Example VII.1.15 *Let*

$$X_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad Y_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

The columns of X_1 and of Y_1 are orthonormal vectors. If we choose unitary matrices

$$U_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad V_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

and

$$Q = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

then we see that

$$QX_1U_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad QY_1V_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus in the space $\mathbb{R}^4$ (or $\mathbb{C}^4$), the canonical angles between the 2-dimensional subspaces spanned by the columns of X_1 and Y_1, respectively, are $\frac{\pi}{4}$, $\frac{\pi}{4}$.

VII.2 The Equation $AX - XB = Y$

We study in some detail the Sylvester equation,

$$AX - XB = Y. \tag{VII.16}$$

Here A is an operator on a Hilbert space $\mathcal{H}$, B is an operator on a Hilbert space $\mathcal{K}$, and X, Y are operators from $\mathcal{K}$ into $\mathcal{H}$. Most of the time we are interested in the situation when $\mathcal{K} = \mathcal{H} = \mathbb{C}^n$, and we will state and prove our results for this special case. The extension to the more general situation is straightforward.

We are given A and B, and we ask the following questions about the above equation. When is there a unique solution X for every Y? What is the form of the solution? Can we estimate $\|X\|$ in terms of $\|Y\|$?

Theorem VII.2.1 *Let A, B be operators with spectra $\sigma(A)$ and $\sigma(B)$, respectively. If $\sigma(A)$ and $\sigma(B)$ are disjoint, then the equation (VII.16) has a unique solution X for every Y.*

Proof. Let $\mathcal{T}$ be the linear operator on the space of operators, defined by $\mathcal{T}(X) = AX - XB$. The conclusion of the theorem can be rephrased as: $\mathcal{T}$ is invertible if $\sigma(A)$ and $\sigma(B)$ are disjoint.

Let $\mathcal{A}(X) = AX$ and $\mathcal{B}(X) = XB$. Then $\mathcal{T} = \mathcal{A} - \mathcal{B}$ and $\mathcal{A}$ and $\mathcal{B}$ commute (regardless of whether A and B do). Hence, $\sigma(\mathcal{T}) \subset \sigma(\mathcal{A}) - \sigma(\mathcal{B})$. If x is an eigenvector of A with eigenvalue α, then the matrix X, one of whose columns is x and the rest of whose columns are zero, is an eigenvector of $\mathcal{A}$ with eigenvalue α. Thus the eigenvalues of $\mathcal{A}$ are just the eigenvalues of A, each counted n times as often. So $\sigma(\mathcal{A}) = \sigma(A)$. In the same way, $\sigma(\mathcal{B}) = \sigma(B)$. Hence $\sigma(\mathcal{T}) \subset \sigma(A) - \sigma(B)$. So, if $\sigma(A)$ and $\sigma(B)$ are disjoint, then $0 \notin \sigma(\mathcal{T})$. Thus, $\mathcal{T}$ is invertible. ■

It is instructive to note that the scalar equation $ax - xb = y$ has a unique solution x for every y if $a - b \neq 0$. The condition $0 \notin \sigma(A) - \sigma(B)$ can be interpreted to be a generalisation of this to the matrix case. This analogy will be helpful in the discussion that follows.

Consider the scalar equation $ax - xb = y$. Exclude the trivial cases in which $a = b$ and in which either a or b is zero. The solution to this equation can be written as

$$x = a^{-1}\left(1 - \frac{b}{a}\right)^{-1} y.$$

If $|b| < |a|$, the middle factor on the right can be expanded as a convergent power series, and we can write

$$x = a^{-1}\sum_{n=0}^{\infty}\left(\frac{b}{a}\right)^{n} y \quad = \sum_{n=0}^{\infty} a^{-n-1} y b^{n}.$$

This is surely a complicated way of writing $x = y/(a - b)$. However, it suggests, in the operator case, the form of the solution given in the theorem below. For the proof of the theorem we will need the **spectral radius formula**. This says that the spectral radius of any operator A is given by the formula

$$\operatorname{spr}(A) = \lim_{n\to\infty} \|A^n\|^{1/n}.$$

Theorem VII.2.2 *Let A, B be operators such that $\sigma(B) \subset \{z : |z| < \rho\}$ and $\sigma(A) \subset \{z : |z| > \rho\}$ for some $\rho > 0$. Then the solution of the equation $AX - XB = Y$ is*

$$X = \sum_{n=0}^{\infty} A^{-n-1}\, Y B^{n}. \tag{VII.17}$$

Proof. We will prove that the series converges. It is then easy to see that X so defined is a solution of the equation.

Choose $\rho_1 < \rho < \rho_2$ such that $\sigma(B)$ is contained in the disk $\{z : |z| < \rho_1\}$ and $\sigma(A)$ is outside the disk $\{z : |z| < \rho_2\}$. Then $\sigma(A^{-1})$ is inside the disk $\{z : |z| < \rho_2^{-1}\}$. By the spectral radius formula, there exists a positive

integer N such that for $n \geq N$, $\|B^n\| \leq \rho_1^n$ and $\|A^{-n}\| < \rho_2^{-n}$. Hence, for $n \geq N$, $\|A^{-n-1}YB^n\| \leq (\rho_1/\rho_2)^n \|A^{-1}Y\|$. Thus the series in (VII.17) is convergent. ■

Another solution to (VII.16) is obtained from the following considerations. If $\text{Re}(b-a) < 0$, the integral $\int_0^\infty e^{t(b-a)}dt$ is convergent and has the value $\frac{1}{a-b}$. Thus, in this case, the solution of the equation $ax - xb = y$ can be expressed as $x = \int_0^\infty e^{t(b-a)} y\, dt$. This is the motivation for the following theorem.

Theorem VII.2.3 *Let A and B be operators whose spectra are contained in the open right half-plane and the open left half-plane, respectively. Then the solution of the equation $AX - XB = Y$ can be expressed as*

$$X = \int_0^\infty e^{-tA} Y e^{tB}\, dt. \tag{VII.18}$$

Proof. It is easy to see that the hypotheses ensure that the integral given above is convergent. If X is the operator defined by this integral, then

$$\begin{aligned} AX - XB &= \int_0^\infty (Ae^{-tA}Ye^{tB} - e^{-tA}Ye^{tB}B)dt \\ &= -e^{-tA}Ye^{tB}\Big|_0^\infty = Y. \end{aligned}$$

So X is indeed the solution of the equation. ■

Notice that in both the theorems above we made a special assumption about the way $\sigma(A)$ and $\sigma(B)$ are separated. No such assumption is made in the theorem below. Once again, it is helpful to consider the scalar case first. Note that

$$\frac{1}{(a-\zeta)(b-\zeta)} = \left(\frac{1}{a-\zeta} - \frac{1}{b-\zeta}\right)\frac{1}{b-a}.$$

So, if Γ is any closed contour in the complex plane with winding numbers 1 around a and 0 around b, then by Cauchy's integral formula we have

$$\int_\Gamma \frac{1}{(a-\zeta)(b-\zeta)} d\zeta = \frac{2\pi\, i}{a-b}.$$

Thus the solution of the equation $ax - xb = y$ can be expressed as

$$x = \frac{1}{2\pi i}\int_\Gamma \frac{y}{(a-\zeta)(b-\zeta)}\, d\zeta.$$

The appropriate generalisation for operators is the following.

Theorem VII.2.4 *Let A and B be operators whose spectra are disjoint from each other. Let Γ be any closed contour in the complex plane with winding numbers 1 around $\sigma(A)$ and 0 around $\sigma(B)$. Then the solution of the equation $AX - XB = Y$ can be expressed as*

$$X = \frac{1}{2\pi i} \int_{\Gamma} (A-\zeta)^{-1} Y (B-\zeta)^{-1} d\zeta. \qquad \text{(VII.19)}$$

Proof. If $AX - XB = Y$, then for every complex number ζ, $(A-\zeta)X - X(B-\zeta) = Y$. If $A-\zeta$ and $B-\zeta$ are invertible, this gives

$$X(B-\zeta)^{-1} - (A-\zeta)^{-1}X = (A-\zeta)^{-1}Y(B-\zeta)^{-1}.$$

Integrate both sides over the given contour Γ and note that $\int_{\Gamma}(B-\zeta)^{-1}d\zeta = 0$ and $-\int_{\Gamma}(A-\zeta)^{-1}d\zeta = 2\pi i I$. This proves the theorem. ∎

Our principal interest is in the case when A and B in the equation (VII.16) are both normal or, even more specially, Hermitian or unitary. In these cases more special forms of the solution can be obtained.

Let A and B be both Hermitian. Then iA and iB are skew-Hermitian, and hence their spectra lie on the imaginary line. This is just the opposite of the situation that Theorem VII.2.3 was addressed to. If we were to imitate that solution, we would try out the integral $\int_0^\infty e^{-itA} Y e^{itB} dt$. This, however, does not converge. This can be remedied by inserting a convergence factor: a function f in $L^1(\mathbb{R})$. If we set

$$X = \int_{-\infty}^{\infty} e^{-itA} Y e^{itB} f(t) dt,$$

then this is a well-defined operator for each $f \in L^1(\mathbb{R})$, since for each t the exponentials occurring above are unitary operators. Of course, such an X need not be a solution of the equation (VII.16). Can a special choice of f make it so? Once again, it is instructive to first examine the scalar case. In this case, the above expression reduces to

$$x = y \hat{f}(a-b),$$

where $\hat{f}$ is the Fourier transform of f, defined as

$$\hat{f}(s) = \int_{-\infty}^{\infty} e^{-its} f(t) dt. \qquad \text{(VII.20)}$$

So, if we choose an f such that $\hat{f}(a-b) = \frac{1}{a-b}$, we do have $ax - xb = y$. The following theorem generalises this to operators.

Theorem VII.2.5 *Let A, B be Hermitian operators whose spectra are disjoint from each other. Let f be any function in $L^1(\mathbb{R})$ such that $\hat{f}(s) = \frac{1}{s}$ whenever $s \in \sigma(A) - \sigma(B)$. Then the solution of the equation $AX - XB = Y$ can be expressed as*

$$X = \int_{-\infty}^{\infty} e^{-itA} Y e^{itB} f(t) dt. \tag{VII.21}$$

Proof. Let α and β be eigenvalues of A and B with eigenvectors u and v, respectively. Then, using the fact that e^{itA} is unitary and its adjoint is e^{-itA}, we see that

$$\begin{aligned} \langle u, Ae^{-itA} Y e^{itB} v \rangle &= \langle e^{itA} Au,\ Y e^{itB} v \rangle \\ &= e^{it(\beta-\alpha)} \alpha \langle u, Yv \rangle. \end{aligned}$$

A similar consideration shows that

$$\langle u, e^{-itA} Y e^{itB} Bv \rangle = e^{it(\beta-\alpha)} \beta \langle u, Yv \rangle.$$

Hence, if X is given by (VII.21), we have

$$\begin{aligned} \langle u, (AX - XB)v \rangle &= (\alpha - \beta) \langle u, Yv \rangle \int_{-\infty}^{\infty} e^{it(\beta-\alpha)} f(t) dt \\ &= (\alpha - \beta) \langle u, Yv \rangle \hat{f}(\alpha - \beta) \\ &= \langle u, Yv \rangle. \end{aligned}$$

Since eigenvectors of A and B both span the whole space, this shows that $AX - XB = Y$. ■

The two theorems below can be proved using the same argument as above. For a function f in $L^1(\mathbb{R}^2)$ we will use the notation $\hat{f}$ for its Fourier transform, defined as

$$\hat{f}(s_1, s_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(t_1 s_1 + t_2 s_2)} f(t_1, t_2) dt_1 dt_2.$$

Theorem VII.2.6 *Let A and B be normal operators whose spectra are disjoint from each other. Let $A = A_1 + iA_2$, $B = B_1 + iB_2$, where A_1 and A_2 are commuting Hermitian operators and so are B_1 and B_2. Let f be any function in $L^1(\mathbb{R}^2)$ such that $\hat{f}(s_1, s_2) = \frac{1}{s_1 + is_2}$ whenever $s_1 + is_2 \in$*

$\sigma(A) - \sigma(B)$. Then the solution of the equation $AX - XB = Y$ can be expressed as

$$X = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(t_1A_1+t_2A_2)}Ye^{i(t_1B_1+t_2B_2)}f(t_1,t_2)dt_1dt_2. \qquad \text{(VII.22)}$$

Theorem VII.2.7 *Let A and B be unitary operators whose spectra are disjoint from each other. Let $\{a_n\}_{-\infty}^{\infty}$ be any sequence in ℓ_1 such that*

$$\sum_{n=-\infty}^{\infty} a_n e^{in\theta} = \frac{1}{1-e^{i\theta}} \quad \text{whenever} \quad e^{i\theta} \in (\sigma(A))^{-1}\cdot\sigma(B).$$

Then the solution of the equation $AX - XB = Y$ can be expressed as

$$X = \sum_{n=-\infty}^{\infty} a_n A^{-n-1}YB^n. \qquad \text{(VII.23)}$$

The different formulae obtained above lead to estimates for $|||X|||$ when A and B are normal. These estimates involve $|||Y|||$ and the separation δ between $\sigma(A)$ and $\sigma(B)$, where

$$\delta = \text{dist}(\sigma(A),\sigma(B)) = \min\{|\lambda-\mu| : \lambda\in\sigma(A),\ \mu\in\sigma(B)\}.$$

The special case of the Frobenius norm is the simplest.

Theorem VII.2.8 *Let A and B be normal matrices, and let $\delta = \text{dist}(\sigma(A),\ \sigma(B)) > 0$. Then the solution X of the equation $AX - XB = Y$ satisfies the inequality*

$$\|X\|_2 \le \frac{1}{\delta}\|Y\|_2. \qquad \text{(VII.24)}$$

Proof. If A and B are both diagonal with diagonal entries $\lambda_1,\ldots,\lambda_n$ and $\mu_1,\ldots,\mu_n$, respectively, then the entries of X and Y are related by the equation $x_{ij} = y_{ij}/(\lambda_i - \mu_j)$. From this (VII.24) follows immediately.

If A, B are any normal matrices, we can find unitary matrices U, V and diagonal matrices A', B' such that $A = UA'U^*$ and $B = VB'V^*$. The equation $AX - XB = Y$ can be rewritten as

$$UA'U^*X - XVB'V^* = Y$$

and then as

$$A'(U^*XV) - (U^*XV)B' = U^*YV.$$

So, we now have the same type of equation but with diagonal A', B'. Hence,

$$\|U^*XV\|_2 \le \frac{1}{\delta}\|U^*YV\|_2.$$

By the unitary invariance of the Frobenius norm this is the same as the inequality (VII.24). ■

Example VII.2.9 *If A or B is not normal, no inequality like (VII.24) is true in general. For example, if $A = Y = I$ and $B = \begin{pmatrix} 0 & t \\ 0 & 0 \end{pmatrix}$, then the equation $AX - XB = Y$ has the solution $X = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$. Here $\delta = 1$, $\|Y\|_2 = \sqrt{2}$, but $\|X\|_2$ can be made arbitrarily large by choosing t large. Thus we cannot even have a bound like $\|X\|_2 \leq \frac{c}{\delta}\|Y\|_2$ for any constant c in this case.*

Example VII.2.10 *In this example all matrices involved are Hermitian:*

$$A = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} -3 & 0 \\ 0 & 1 \end{pmatrix},$$

$$X = \begin{pmatrix} 1 & \sqrt{15} \\ \sqrt{15} & 3 \end{pmatrix}, \quad Y = \begin{pmatrix} 6 & 2\sqrt{15} \\ 2\sqrt{15} & -6 \end{pmatrix}.$$

Then $AX - XB = Y$. Here $\delta = 2$. But, for the operator norm, $\|X\| > \frac{1}{2}\|Y\|$. Thus, the inequality $\|X\| \leq \frac{1}{\delta}\|Y\|$ need not hold even for Hermitian A, B.

In the next theorems we will see that we do have $|||X||| \leq \frac{c}{\delta}|||Y|||$ for a small constant c when A and B are normal. When the spectra of A and B are separated in a special way, we can choose $c = 1$.

Theorem VII.2.11 *Let A and B be normal operators such that the spectrum of B is contained in a disk $D(a, \rho)$ and the spectrum of A lies outside a concentric disk $D(a, \rho+\delta)$. Then, the solution of the equation $AX - XB = Y$ satisfies the inequality*

$$|||X||| \leq \frac{1}{\delta}|||Y||| \tag{VII.25}$$

for every unitarily invariant norm.

Proof. Applying a translation, we can assume that $a = 0$. Then the solution X can be expressed as the infinite series (VII.17). From this we get

$$\begin{aligned} |||X||| &\leq \sum_{n=0}^{\infty} \|A^{-1}\|^{n+1} |||Y|||\ \|B\|^n \\ &\leq |||Y||| \sum_{n=0}^{\infty} (\rho + \delta)^{-n-1}\ \rho^n \\ &= \frac{1}{\delta}|||Y|||. \end{aligned}$$

■

Either by taking a limit $\rho \to \infty$ in the above argument or by using the form of the solution (VII.18), we can prove the following.

Theorem VII.2.12 *Let A and B be normal operators with $\sigma(A)$ and $\sigma(B)$ lying in half-planes separated by a strip of width δ. Then the solution of the equation $AX - XB = Y$ satisfies the inequality (VII.25).*

Exercise VII.2.13 *Here is an alternate proof of Theorem VII.2.11. Assume, without loss of generality, that $a = 0$. Then A is invertible. Write $X = A^{-1}(Y + XB)$ and obtain the inequality (VII.25) directly from this.*

Exercise VII.2.14 *Choose unit vectors u and v such that $Xv = \|X\|u$ and $X^*u = \|X\|v$; i.e., u and v are left and right singular vectors of X corresponding to its largest singular value. Then $\langle u, (AX - XB)v\rangle = \|X\|(\langle u, Au\rangle - \langle v, Bv\rangle)$. Use this to prove Theorem VII.2.11 in the special case of the operator norm.*

Theorem VII.2.15 *Let A and B be Hermitian operators with* dist $(\sigma(A), \sigma(B)) = \delta > 0$. *Then, the solution of the equation $AX - XB = Y$ satisfies the inequality*

$$|||X||| \leq \frac{c_1}{\delta}\, |||Y||| \tag{VII.26}$$

for every unitarily invariant norm, where c_1 is a positive real number defined as

$$c_1 = \inf\{\|f\|_{L^1} : f \in L^1(\mathbb{R}),\ \hat{f}(s) = \frac{1}{s} \text{ when } |s| \geq 1\}. \tag{VII.27}$$

Proof. Let f_δ be any function in $L^1(\mathbb{R})$ such that $\hat{f}_\delta(s) = \frac{1}{s}$ whenever $|s| \geq \delta$. By Theorem VII.2.5 we have

$$X = \int_{-\infty}^{\infty} e^{-itA} Y e^{itB} f_\delta(t) dt.$$

Hence,

$$|||X||| \leq |||Y||| \int_{-\infty}^{\infty} |f_\delta(t)| dt = \frac{1}{\delta} |||Y||| \int_{-\infty}^{\infty} |f(t)| dt,$$

where $f(t) = f_\delta(t/\delta)$. Note that $\hat{f}(s) = \frac{1}{s}$ whenever $|s| \geq 1$. Any f with this property satisfies the above inequality. ■

Exactly the same argument, using Theorem VII.2.6 now leads to the following.

Theorem VII.2.16 *Let A and B be normal operators with* dist $(\sigma(A), \sigma(B)) = \delta > 0$. *Then the solution of the equation $AX - XB = Y$ satisfies the inequality*

$$|||X||| \leq \frac{c_2}{\delta} |||Y||| \tag{VII.28}$$

for every unitarily invariant norm, where

$$c_2 = \inf\{\|f\|_{L^1} : f \in L^1(\mathbb{R}^2),\ \hat{f}(s_1, s_2) = \frac{1}{s_1 + is_2} \text{ when } s_1^2 + s_2^2 \geq 1\}. \tag{VII.29}$$

The exact evaluation of the constants c_1 and c_2 is an intricate problem in Fourier analysis. This is discussed in the Appendix at the end of this chapter. It is known that

$$c_1 = \frac{\pi}{2}$$

and

$$c_2 \leq \frac{\pi}{2} \int_0^\pi \frac{\sin t}{t} dt < 2.91.$$

Further, with this value of c_1, the inequality (VII.26) is sharp.

VII.3 Perturbation of Eigenspaces

Given a normal operator A and a subset S of $\mathbb{C}$, we will write $P_A(S)$ for the orthogonal projection onto the subspace spanned by the eigenvectors of A corresponding to those of its eigenvalues that lie in S.

If S_1 and S_2 are two disjoint sets, and if $E = P_A(S_1)$ and $F = P_A(S_2)$, then E and F are mutually orthogonal. If A and B are two normal operators, and if $E = P_A(S_1)$ and $F = P_B(S_2)$, then we might expect that if B is close to A and S_1 and S_2 are far apart, then E is nearly orthogonal to F. This is made precise in the theorems below.

Theorem VII.3.1 *Let A, B be normal operators. Let S_1 and S_2 be two subsets of the complex plane that are separated by either an annulus of width δ or a strip of width δ. Let $E = P_A(S_1)$, $F = P_B(S_2)$. Then, for every unitarily invariant norm,*

$$|||EF||| \leq \frac{1}{\delta}|||E(A-B)F||| \leq \frac{1}{\delta}|||A-B|||. \tag{VII.30}$$

Proof. Since E commutes with A and F with B, the first inequality in (VII.30) can be written as

$$|||EF||| \leq \frac{1}{\delta}|||AEF - EFB|||.$$

Now let $EF = X$. This is an operator from the space $\operatorname{ran} F$ to the space $\operatorname{ran} E$. Restricted to these spaces, the operators B and A have their spectra inside S_2 and S_1, respectively. Thus the above inequality follows from Theorem VII.2.11 when S_1 and S_2 are separated by an annulus, and from Theorem VII.2.12 when they are separated by a strip.

The second inequality in (VII.30) is true because $\|E\| = \|F\| = 1$. ■

The special case of this theorem when A and B are Hermitian, S_1 is an interval $[a, b]$ and S_2 the complement in $\mathbb{R}$ of the interval $(a - \delta,\ b + \delta)$, is known as the **Davis-Kahan sinΘ Theorem**. (We saw in Section 1 that $\|EF\|$ is the sine of the angle between $\operatorname{ran} E$ and $\operatorname{ran} F^{\perp}$.)

With no special assumption on the way S_1 and S_2 are separated, we can derive the following two theorems from Theorems VII.2.15 and VII.2.16 by the argument used above.

Theorem VII.3.2 *Let A and B be Hermitian operators, and let S_1, S_2 be any two subsets of $\mathbb{R}$ such that* dist $(S_1, S_2) = \delta > 0$. *Let $E = P_A(S_1)$, $F = P_B(S_2)$. Then, for every unitarily invariant norm,*

$$|||EF||| \leq \frac{c_1}{\delta} |||E(A - B)F||| \leq \frac{c_1}{\delta} |||A - B|||, \tag{VII.31}$$

where c_1 is the constant defined by (VII.27). (We know that $c_1 = \frac{\pi}{2}$.)

Theorem VII.3.3 *Let A and B be normal operators, and let S_1, S_2 be any two subsets of the complex plane such that* dist $(S_1, S_2) = \delta > 0$. *Let $E = P_A(S_1)$, $F = P_B(S_2)$. Then, for every unitarily invariant norm,*

$$|||EF||| \leq \frac{c_2}{\delta} |||E(A - B)F||| \leq \frac{c_2}{\delta} |||A - B|||, \tag{VII.32}$$

where c_2 is the constant defined by (VII.29). (We know that $c_2 < 2.91$.)

Finally, note that for the Frobenius norm alone, we have a stronger result as a consequence of Theorem VII.2.8.

Theorem VII.3.4 *Let A and B be normal operators and let S_1, S_2 be any two subsets of the complex plane such that* dist $(S_1, S_2) = \delta > 0$. *Let $E = P_A(S_1)$, $F = P_B(S_2)$. Then*

$$\|EF\|_2 \leq \frac{1}{\delta} \|E(A - B)F\|_2 \leq \frac{1}{\delta} \|A - B\|_2.$$

VII.4 A Perturbation Bound for Eigenvalues

An important corollary of Theorem VII.3.3 is the following bound for the distance between the eigenvalues of two normal matrices.

Theorem VII.4.1 *There exists a constant $c, 1 < c < 3$, such that the optimal matching distance $d(\sigma(A),\ \sigma(B))$ between the eigenvalues of any two normal matrices A and B is bounded as*

$$d(\sigma(A),\ \sigma(B)) \leq c\|A - B\|. \tag{VII.33}$$

Proof. We will show that the inequality (VII.33) is true if $c = c_2$, the constant in the inequality (VII.32).

Let $\eta = c_2\|A - B\|$ and suppose $d(\sigma(A),\ \sigma(B)) > \eta$. Then we can find a $\delta > \eta$ such that $d(\sigma(A),\ \sigma(B)) > \delta$. By the Marriage Theorem, this is possible if and only if there exists a set S_1 consisting of k eigenvalues of $A, 1 \le k \le n$, such that the δ-neighbourhood $\{z : \text{dist}(z, S_1) \le \delta\}$ contains less than k eigenvalues of B. Let S_2 be the set of all eigenvalues of B outside this neighbourhood. Then $\text{dist}(S_1, S_2) \ge \delta$. Let $E = P_A(S_1),\ F = P_B(S_2)$. Then the dimension of the range of E is k, and that of the range of F is at least $n - k + 1$. Hence $\|EF\| = 1$. On the other hand, the inequality (VII.32) implies that

$$\|EF\| \le \frac{c_2}{\delta}\ \|A - B\| = \frac{\eta}{\delta} < 1.$$

This is a contradiction. So the inequality (VII.33) is valid if we choose $c = c_2(< 2.91)$.

Example VI.3.13 shows that any constant c for which the inequality (VII.33) is valid for all normal matrices A, B must be larger than 1.018. ■

We should remark that, for Hermitian matrices, this reasoning using Theorem VII.3.2 will give the inequality $d(\sigma(A),\ \sigma(B)) \le \frac{\pi}{2}\|A - B\|$. However, in this case, we have the stronger inequality $d(\sigma(A),\ \sigma(B)) \le \|A - B\|$. So, this may not be the best method of deriving spectral variation bounds. However, for normal matrices, nothing more effective has been found yet.

VII.5 Perturbation of the Polar Factors

Let $A = UP$ be the polar decomposition of A. The positive part P in this decomposition is $P = |A| = (A^*A)^{1/2}$ and is always unique. The unitary part U is unique if A is invertible. Then $U = AP^{-1}$.

It is of interest to know how a change in A affects its polar factors U and P. Some results on this are proved below.

Let A and B be invertible operators with polar decompositions $A = UP$ and $B = VQ$, respectively, where U and V are unitary, and P and Q are positive. Then,

$$|||A - B||| = |||UP - VQ||| = |||P - U^*VQ|||$$

for every unitarily invariant norm. By symmetry,

$$|||A - B||| = |||Q - V^*UP|||.$$

Let

$$Y = P - U^*VQ, \quad Z = Q - V^*UP.$$

Then

$$Y + Z^* = P(I - U^*V) + (I - U^*V)Q. \tag{VII.34}$$

This equation is of the form studied in Section VII.2. Note that $\sigma(P)$ is a subset of the real line bounded below by $s_n(A) = \|A^{-1}\|^{-1}$ and $\sigma(Q)$ is a subset of the real line bounded below by $s_n(B) = \|B^{-1}\|^{-1}$. Hence, dist $(\sigma(P), \sigma(-Q)) = s_n(A) + s_n(B)$. Hence, by Theorem VII.2.11,

$$|||I - U^*V||| \leq \frac{1}{s_n(A) + s_n(B)} |||Y + Z^*|||.$$

Since $|||Y||| = |||Z||| = |||A - B|||$ and $|||I - U^*V||| = |||U - V|||$, this gives the following theorem.

Theorem VII.5.1 *Let A and B be invertible operators, and let U, V be the unitary factors in their polar decompositions. Then*

$$|||U - V||| \leq \frac{2}{\|A^{-1}\|^{-1} + \|B^{-1}\|^{-1}} |||A - B||| \tag{VII.35}$$

for every unitarily invariant norm $|||\cdot|||$.

Exercise VII.5.2 *Find matrices A, B for which (VII.35) is an equality.*

Exercise VII.5.3 *Let A, B be invertible operators. Show that*

$$|||\, |A| - |B| \,||| \leq \left(1 + \frac{2m}{\|A^{-1}\|^{-1} + \|B^{-1}\|^{-1}}\right) |||A - B|||, \tag{VII.36}$$

where $m = \min(\|A\|, \|B\|)$.

For the Frobenius norm alone, a simpler inequality can be obtained as shown below.

Lemma VII.5.4 *Let f be a Lipschitz continuous function on $\mathbb{C}$ satisfying the inequality*

$$|f(z) - f(w)| \leq k|z - w|, \quad \textit{for all } z, w \in \mathbb{C}.$$

Then, for all matrices X and all normal matrices A, we have

$$\|f(A)X - Xf(A)\|_2 \leq k\|AX - XA\|_2.$$

Proof. Assume, without loss of generality, that $A = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$. Then, if X is any matrix with entries x_{ij}, we have

$$\begin{aligned} \|f(A)X - Xf(A)\|_2^2 &= \sum_{i,j} |[f(\lambda_i) - f(\lambda_j)]x_{ij}|^2 \\ &\leq k^2 \sum_{i,j} |\lambda_i - \lambda_j|^2 \, |x_{ij}|^2 \\ &= k^2 \|AX - XA\|_2^2. \end{aligned}$$

■

Lemma VII.5.5 *Let f be a function satisfying the conditions of Lemma VII.5.4. Let A, B be any two normal matrices. Then, for every matrix X,*

$$\|f(A)X - Xf(B)\|_2 \leq k\|AX - XB\|_2.$$

Proof. Let $T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$, $Y = \begin{pmatrix} 0 & X \\ 0 & 0 \end{pmatrix}$. Replace A and X in Lemma VII.5.4 by T and Y, respectively. ■

Corollary VII.5.6 *If A and B are normal matrices, then*

$$\| \, |A| - |B| \, \|_2 \leq \|A - B\|_2. \tag{VII.37}$$

Theorem VII.5.7 *Let A, B be any two matrices. Then*

$$\| \, |A| - |B| \, \|_2^2 + \| \, |A^*| - |B^*| \, \|_2^2 \leq 2\|A - B\|_2^2. \tag{VII.38}$$

Proof. Let $T = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}$, $S = \begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix}$. Then T and S are Hermitian. Note that $|T| = \begin{pmatrix} |A^*| & 0 \\ 0 & |A| \end{pmatrix}$. So, the inequality (VII.38) follows from (VII.37). ■

It follows from (VII.38) that

$$\| \, |A| - |B| \, \|_2 \leq \sqrt{2} \, \|A - B\|_2. \tag{VII.39}$$

The next example shows that the Lipschitz constant $\sqrt{2}$ in the above inequality cannot be replaced by a smaller number.

Example VII.5.8 *Let*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & \varepsilon \\ 0 & 0 \end{pmatrix}.$$

Then $|A| = A$ and

$$|B| = \frac{1}{\sqrt{1+\varepsilon^2}} \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & \varepsilon^2 \end{pmatrix}.$$

As $\varepsilon \to 0$, the ratio $\frac{\| \, |A| - |B| \, \|_2}{\|A - B\|_2}$ approaches $\sqrt{2}$.

We will continue the study of perturbation of the function $|A|$ in later chapters.

A useful consequence of Theorem VII.5.7 is the following perturbation bound for singular vectors.

Theorem VII.5.9 *Let S_1, S_2 be two subsets of the positive half-line such that $\text{dist}(S_1, S_2) = \delta > 0$. Let A and B be any two matrices. Let E and E' be the orthogonal projections onto the subspaces spanned by the right and the left singular vectors of A corresponding to its singular values in S_1. Let F and F' be the projections associated with B in the same way, corresponding to its singular values in S_2. Then*

$$(\|EF\|_2^2 + \|E'F'\|_2^2)^{1/2} \leq \frac{\sqrt{2}}{\delta} \, \|A - B\|_2. \tag{VII.40}$$

Proof. By Theorem VII.3.4 we have

$$\begin{aligned}\|EF\|_2 &\leq \frac{1}{\delta}\| \, |A| - |B| \, \|_2, \\ \|E'F'\|_2 &\leq \frac{1}{\delta}\| \, |A^*| - |B^*| \, \|_2.\end{aligned}$$

These inequalities, together with (VII.38), lead to the inequality (VII.40). ∎

VII.6 Appendix: Evaluating the (Fourier) constants

The analysis in Section VII.2 has led to some extremal problems in Fourier analysis. Here we indicate how the constants c_1 and c_2 defined by (VII.27) and (VII.29) may be evaluated.

The symbol $\|f\|_1$ will denote the norm in the space L^1 for functions defined on $\mathbb{R}$ or on $\mathbb{R}^2$.

We are required to find a function f in L^1, with minimal norm, such that $\hat{f}(s) = \frac{1}{s}$ when $|s| \geq 1$. Since $\hat{f}$ must be continuous, we might begin by taking a continuous function that coincides with $\frac{1}{s}$ for $|s| \geq 1$ and then taking f to be its inverse Fourier transform. The difficulty is that the function $\frac{1}{s}$ is not in L^1, and hence its inverse Fourier transform may not be defined. Note, however, that the function $\frac{1}{s}$ is square integrable at ∞. So it is the Fourier transform of an L^2 function. We will show that under suitable conditions its inverse Fourier transform is in L^1, and find one that has the least norm.

Since the function $\frac{1}{s}$ is an odd function, it would seem economical to extend it inside the domain $(-1, 1)$, so that the extended function is an odd function on $\mathbb{R}$. This is indeed so. Let $f \in L^1(\mathbb{R})$ and suppose $\hat{f}(s) = \frac{1}{s}$ when $|s| \geq 1$. Let f_{odd} be the odd part of f, $f_{\text{odd}}(t) = \frac{f(t)-f(-t)}{2}$. Then $\hat{f}_{\text{odd}}(s) = \frac{1}{s}$ when $|s| \geq 1$ and $\|f_{\text{odd}}\|_1 \leq \|f\|_1$. Thus the constant c_1 is also the infimum of $\|f\|_1$ over all odd functions in L^1 for which $\hat{f}(s) = \frac{1}{s}$ when $|s| \geq 1$.

Now note that if f is odd, then

$$\begin{aligned}\hat{f}(s) &= \int_{-\infty}^{\infty} f(t)e^{-its}dt = -i\int_{-\infty}^{\infty} f(t) \, \sin ts \, dt \\ &= -i\int_{-\infty}^{\infty} \text{Re}\, f(t) \, \sin ts \, dt + \int_{-\infty}^{\infty} \text{Im}\, f(t) \, \sin ts \, dt.\end{aligned}$$

If this is to be equal to $\frac{1}{s}$ when $|s| \geq 1$, the Fourier transform of $\text{Re} f$ should have its support in $(-1, 1)$. Thus, it is enough to consider purely imaginary

functions f in our extremal problem. Equivalently, let $\mathcal{C}$ be the class of all odd real functions in $L^1(\mathbb{R})$ such that $\hat{f}(s) = \frac{1}{is}$ when $|s| \geq 1$. Then

$$c_1 = \inf\{\|f\|_1 : f \in \mathcal{C}\}. \tag{VII.41}$$

Now, let g be any bounded function with period 2π having a Fourier series expansion $\sum\limits_{n\neq 0} \alpha_n e^{int}$. This last condition means that $\int\limits_0^{2\pi} g(t)dt = 0$. Then, for any f in $\mathcal{C}$,

$$\int\limits_{-\infty}^{\infty} f(t)g(x-t)dt = \sum_{n\neq 0} \frac{\alpha_n}{in} e^{inx}. \tag{VII.42}$$

Note that this expression does not depend on the choice of f.

For a real number x, let sgn$x$ be defined as -1 if x is negative and 1 if x is nonnegative. Let f_0 be an element of $\mathcal{C}$ such that

$$\operatorname{sgn} f_0(t) = \operatorname{sgn}\, \sin t. \tag{VII.43}$$

Note that the function sgn f_0 then satisfies the requirements made on g in the preceding paragraph. Hence, we have

$$\begin{aligned}\int\limits_{-\infty}^{\infty} |f_0(t)|dt &= \int\limits_{-\infty}^{\infty} f_0(t)\operatorname{sgn} f_0(t)dt = -\int\limits_{-\infty}^{\infty} f_0(t)\operatorname{sgn} f_0(-t)dt \\ &= -\int\limits_{-\infty}^{\infty} f(t)\operatorname{sgn} f_0(-t)dt \leq \int\limits_{-\infty}^{\infty} |f(t)|dt\end{aligned}$$

for every $f \in \mathcal{C}$. (Use (VII.42) with $x = 0$ and see the remark following it.) Thus $c_1 = \|f_0\|_1$, where f_0 is any function in $\mathcal{C}$ satisfying (VII.43). We will now exhibit such a function.

We have remarked earlier that it is natural to obtain f_0 as the inverse Fourier transform of a continuous odd function φ such that $\varphi(s) = \frac{1}{s}$ for $|s| \geq 1$. First we must find a good sufficient condition on φ so that its inverse Fourier transform

$$\check{\varphi}(t) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \varphi(s)\ \sin ts\ ds$$

(which, by definition, is in L^2) is in L^1. Suppose φ is differentiable on the whole real line and its derivative φ' is of bounded variation. Then, an integration by parts shows

$$\int\limits_{-\infty}^{\infty} \varphi(s)\ \sin ts\ ds = \frac{1}{t} \int\limits_{-\infty}^{\infty} \cos ts\ \varphi'(s)ds.$$

Another integration by parts, this time for Riemann-Stieltjes integrals, shows that

$$\int_{-\infty}^{\infty} \varphi(s) \sin ts \, ds = \frac{1}{t^2} \int_{-\infty}^{\infty} \sin ts \, d\varphi'(s).$$

As $t \to \infty$ this decays as $\frac{1}{t^2}$. So $\check{\varphi}(t)$ is integrable at ∞. Since $\check{\varphi}$ is in L^2, it is integrable over any bounded interval. Hence, $\check{\varphi}$ is in L^1.

We will now find a function φ_0 that satisfies the conditions of the above paragraph, and show that if $f_0 = \check{\varphi}_0$, then f_0 satisfies the condition (VII.43). One such function is

$$\varphi_0(s) = \begin{cases} 1/s & \text{for } |s| \geq 1 \\ \frac{1}{s} - \frac{\pi}{2} \cot \frac{\pi}{2} s & \text{for } 0 < |s| \leq 1 \\ 0 & \text{for } s = 0. \end{cases} \tag{VII.44}$$

From the familiar series expansion

$$\frac{\pi}{2} \cot \frac{\pi}{2} z = \frac{1}{z} + \sum_{n=1}^{\infty} \frac{2z}{z^2 - 4n^2}$$

(see L.V. Ahlfors, *Complex Analysis*, 2nd ed., p. 188) one sees that

$$\varphi_0(s) = \sum_{n=1}^{\infty} \frac{2s}{4n^2 - s^2} \quad \text{for } 0 < s < 1.$$

This shows that φ_0 is a convex function in $0 < s < 1$, and hence φ_0' is of bounded variation in this domain. On the rest of the positive half-line too φ_0' is of bounded variation. So φ_0 does meet the conditions that are sufficient to ensure that $f_0 = \check{\varphi}_0$ is in $\mathcal{C}$.

Using the definition of φ_0, it is straightforward to verify that for $t > 0$,

$$\begin{aligned} 2f_0(t) &= 1 - \int_0^1 \cot \frac{\pi}{2} s \sin ts \, ds, \\ 2f_0(t) - 2f_0(t+\pi) &= \frac{\sin t}{t} + \frac{\sin(t+\pi)}{t+\pi}, \\ f_0(t) - f_0(t+2\pi) &= \left[\frac{1}{2t} - \frac{1}{t+\pi} + \frac{1}{2(t+2\pi)} \right] \sin t. \end{aligned}$$

The quantity inside the brackets is positive for all t. Since $f_0(t) \to 0$ as $t \to \infty$, we can write

$$\begin{aligned} f_0(t) &= \sum_{n=0}^{\infty} [f_0(t+2n\pi) - f_0(t+(2n+1)\pi)] \sin t \\ &= h(t) \sin t, \end{aligned}$$

where $h(t) > 0$. This shows that f_0 satisfies the condition (VII.43).

Finally, $\|f_0\|_1$ can be evaluated from the available data. We can easily see that

$$\begin{aligned} \text{sgn}\,\sin t &= \frac{4}{\pi}(\sin t + \frac{1}{3}\,\sin 3t + \frac{1}{5}\,\sin 5t + \cdots) \\ &= \frac{2}{\pi i}\sum_{n\,\text{odd}} \frac{1}{n}\,e^{int}. \end{aligned}$$

Hence, using (VII.42) we obtain

$$\begin{aligned} \int_{-\infty}^{\infty} |f_0(t)|dt &= -\int_{-\infty}^{\infty} f_0(t)\,\text{sgn}\,f_0(-t)dt \\ &= \frac{2}{\pi}\sum_{n\text{ odd}} \frac{1}{n^2} = \frac{\pi}{2}. \end{aligned}$$

We have shown that $c_1 = \frac{\pi}{2}$. This result, and its proof given above, are due to B. Sz.-Nagy.

The two-variable problem, through which c_2 is defined, is more complicated. The exact value of c_2 is not known. We will show that c_2 is finite by showing that there does exist a function f in $L^1(\mathbb{R}^2)$ such that $\hat{f}(s_1, s_2) = \frac{1}{s_1+is_2}$ when $s_1^2 + s_2^2 \geq 1$. We will then sketch an argument that leads to an estimate of c_2, skipping the technical details.

It is convenient to identify a point (x, y) in $\mathbb{R}^2$ with the complex variable $z = x + iy$. The differential operator $\frac{d}{d\bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\,\frac{\partial}{\partial y}\right)$ annihilates every complex holomorphic function. It is a well-known fact (see, e.g., W. Rudin, *Functional Analysis*, p. 205) that the Fourier transform of the tempered distribution $\frac{1}{z}$ is $-\frac{2\pi i}{z}$. (The normalisations we have chosen are different from those of Rudin.)

Let φ be a C^∞ function on $\mathbb{R}^2$ that vanishes in a neighbourhood of the origin, and is 1 outside another neighbourhood of the origin. Let $\psi(z) = \frac{\varphi(z)}{z}$. We will show that the inverse Fourier transform $\check{\psi}$ is in L^1. Note that

$$\eta(z) := \frac{d}{d\bar{z}}\psi(z) = \frac{1}{z}\,\frac{d\varphi(z)}{d\bar{z}}.$$

This is a C^∞ function with compact support. Hence, η is in the Schwartz space $\mathcal{S}$. Let $\check{\eta} \in \mathcal{S}$ be its inverse Fourier transform. Then (ignoring constant factors) $\check{\psi}(z) = \check{\eta}(z)/z$. Since $\check{\eta}$ is integrable at ∞, so is $\check{\psi}$. At the origin, $\frac{1}{z}$ is integrable and $\check{\eta}(z)$ bounded. Hence $\check{\psi}$ is integrable at the origin.

This shows that $c_2 < \infty$.

Consider the tempered distribution $f_0(z) = \frac{-1}{2\pi i z}$. We know that $\hat{f}_0(\xi) = \frac{1}{\xi}$. However, $f_0 \notin L^1$. To fix it up we seek an element p in the space of tempered distributions $\mathcal{S}'$ such that

(i) $\hat{p} \in L^1$ and supp $\hat{p}$ is contained in the unit disk D,

(ii) if $f = f_0 + p$, then f is in L^1 .

Note that $c_2 = \inf \|f\|_1$ over such p.

Writing $z = re^{i\theta}$, one sees that

$$\|f\|_1 = \frac{1}{2\pi} \int_0^\infty r dr \int_{-\pi}^{\pi} |\frac{1}{r} - ie^{i\theta} 2\pi p(z)| d\theta.$$

Let

$$F(r) = \int_{-\pi}^{\pi} ie^{i\theta} p(z) d\theta. \tag{VII.45}$$

Then

$$\int_{-\pi}^{\pi} |\frac{1}{r} - ie^{i\theta} 2\pi p(z)| d\theta \geq 2\pi |\frac{1}{r} - F(r)|,$$

and there is equality here if $e^{i\theta} p(re^{i\theta})$ is independent of θ. Hence, we can restrict attention to only those p that satisfy the additional condition

(iii) $zp(z)$ is a radial function.

Putting

$$G(r) = 1 - rF(r), \tag{VII.46}$$

we see that

$$c_2 = \inf \int_0^\infty |G(r)| dr, \tag{VII.47}$$

where G is defined via (VII.45) and (VII.46) for all p that satisfy the conditions (i), (ii), and (iii) above. The two-variable minimisation problem is thus reduced to a one-variable problem.

Using the conditions on p, one can characterise the functions G that enter here. This involves a little more intricate analysis, which we will skip. The conclusion is that the functions G that enter in (VII.47) are all L^1 functions of the form $G = \hat{g}$, where g is a continuous even function supported in $[-1, 1]$ such that $\int_{-1}^{1} g(t) dt = 1$. In other words,

$$c_2 = \inf \{ \int_0^\infty |\hat{g}(t)| dt : g \text{ even, supp } g = [-1, 1], \int g = 1, \hat{g} \in L^1 \}. \tag{VII.48}$$

If we choose g to be the function $g(t) = 1 - |t|$, then $\hat{g}(t) = \sin^2(\frac{t}{2})(\frac{t}{2})^2$. This gives the estimate $c_2 \leq \pi$.

A better estimate is obtained from the function

$$g(t) = \frac{\pi}{4} \cos \frac{\pi}{2} t \quad \text{for} \quad |t| \leq 1.$$

Then

$$\hat{g}(t) = \pi^2 \frac{\cos t}{\pi^2 - 4t^2}.$$

A little computation shows that

$$\int_0^\infty |\hat{g}(t)| dt = \frac{\pi}{2} \int_0^\pi \frac{\sin t}{t} dt < 2.90901.$$

Thus $c_2 < 2.91$.

The interested reader may find the details in the paper *An extremal problem in Fourier analysis with applications to operator theory,* by R. Bhatia, C. Davis, and P. Koosis, J. Functional Analysis, 82 (1989) 138-150.

VII.7 Problems

Problem VII.6.1. Let $\mathcal{E}$ be any subspace of $\mathbb{C}^n$. For any vector x let

$$\delta(x, \mathcal{E}) = \min_{y \in \mathcal{E}} \|x - y\|.$$

Then $\delta(x, \mathcal{E})$ is equal to $\|(I - E)x\|$. If $\mathcal{E}, \mathcal{F}$ are two subspaces of $\mathbb{C}^n$, let

$$\rho(\mathcal{E}, \mathcal{F}) = \max\{ \max_{\substack{x \in \mathcal{E} \\ \|x\|=1}} \delta(x, \mathcal{F}), \max_{\substack{y \in \mathcal{F} \\ \|y\|=1}} \delta(y, \mathcal{F})\}.$$

Let dim $\mathcal{E}$ = dim $\mathcal{F}$, and let Θ be the angle operator between $\mathcal{E}$ and $\mathcal{F}$. Show that

$$\rho(\mathcal{E}, \mathcal{F}) = \|\sin \Theta\| = \|E - F\|,$$

where E and F are the orthogonal projections onto the spaces $\mathcal{E}$ and $\mathcal{F}$.

Problem VII.6.2. Let A, B be operators whose spectra are disjoint from each other. Show that the operator $\begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$ is similar to $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ for every C.

Problem VII.6.3. Let A, B be operators whose spectra are disjoint from each other. Show that if C commutes with $A + B$ and with AB, then C commutes with both A and B.

Problem VII.6.4. The equation $AX + XA^* = -I$ is called the **Lyapunov equation**. Show that if $\sigma(A)$ is contained in the open left half-plane, then the Lyapunov equation has a unique solution X, and this solution is positive and invertible.

Problem VII.6.5. Let A and B be any two matrices. Suppose that all singular values of A are at a distance greater than δ from any singular value of B. Show that for every X,

$$\|X\|_2 \le \frac{1}{\delta}\left(\frac{\|AX - XB\|_2^2 + \|A^*X - XB^*\|_2^2}{2}\right)^{1/2}.$$

Problem VII.6.6. Let A, B be normal operators. Let S_1 and S_2 be two subsets of the complex plane separated by a strip of width δ. Let $E = P_A(S_1)$, $F = P_B(S_2)$. Suppose $E(A-B)E = 0$. If $T(t)$ is the function $T(t) = t/\sqrt{1-t^2}$, show that

$$\|T(|EF|)\| \le \frac{1}{\delta}\|E(A-B)\|.$$

Prove that this inequality is also valid for all unitarily invariant norms. This is called the **tanΘ theorem.**

Problem VII.6.7. Show that the inequality (VII.28) cannot be true if $c_2 < \frac{\pi}{2}$. (Hint: Choose the trace norm and find suitable unitary matrices A, B.)

Problem VII.6.8. Show that the conclusion of Theorem VII.2.8 cannot be true for any Schatten p-norm if $p \neq 2$.

Problem VII.6.9. Let A, B be unitary matrices, and let dist $(\sigma(A), \sigma(B)) = \delta = \sqrt{2}$. If some eigenvalue of A is at distance greater than $\sqrt{2}$ from $\sigma(B)$, then $\sigma(A)$ and $\sigma(B)$ can be separated by a strip of width $\sqrt{2}$. In this case, the solution of $AX - XB = Y$ can be obtained from Theorem VII.2.3. Assume that all points of $\sigma(A)$ are at distance $\sqrt{2}$ from all points of $\sigma(B)$. Show that the solution one obtains using Theorem VII.2.7 in this case is

$$X = 1/2\, A^{-1}Y - 1/4\, YB + 1/4\, A^{-2}YB^{-1}.$$

If $\sigma(A) = \{1, -1\}$ and $\sigma(B) = \{i, -i\}$, this reduces to

$$X = 1/2\,(AY - YB).$$

Problem VII.6.10. A reformulation of the Sylvester equation in terms of tensor products is outlined below. Let φ be the natural isomorphism between the Hilbert spaces $\mathcal{H}\otimes\mathcal{H}^*$ and $\mathcal{L}(\mathcal{H})$ constructed in Exercise I.4.4. Show that for every operator A and for each E_{ij},

$$\begin{aligned} \varphi(A\otimes I)\varphi^{-1}(E_{ij}) &= AE_{ij}, \\ \varphi(I\otimes A)\varphi^{-1}(E_{ij}) &= E_{ij}A^T, \end{aligned}$$

where A^T is the transpose of A.

Thus the multiplication operator $\mathcal{A}(X) = AX$ on $\mathcal{L}(\mathcal{H})$ can be identified with the operator $A \otimes I$ on $\mathcal{H} \otimes \mathcal{H}^*$, and the operator $\mathcal{B}(X) = XB$ can be identified with $I \otimes B^T$. The operator $\mathcal{T} = \mathcal{A} - \mathcal{B}$ then corresponds to $A \otimes I - I \otimes B^T$.

Use this to give another proof of Theorem VII.2.1.

Sometimes it is more convenient to identify $\mathcal{L}(\mathcal{H})$ with $\mathcal{H} \otimes \mathcal{H}$ instead of $\mathcal{H} \otimes \mathcal{H}^*$. In this case, we have a bijection φ from $\mathcal{H} \otimes \mathcal{H}$ onto $\mathcal{L}(\mathcal{H})$, that is linear in the first variable and conjugate-linear in the second. With this identification, the operator $\mathcal{A}$ on $\mathcal{L}(\mathcal{H})$ corresponds to the operator $A \otimes I$, while the operator $\mathcal{B}$ corresponds to $I \otimes B^*$.

VII.8 Notes and References

Angles between two subspaces of $\mathbb{R}^n$ were studied in detail by C. Jordan, *Essai sur la géométrie à n dimensions*, Bull. Soc. Math. France, 3 (1875) 103-174. These ideas were reinvented, developed, and used by several mathematicians and statisticians. The detailed analysis of these in connection with perturbation of eigenvectors was taken up by C. Davis, *Separation of two linear subspaces*, Acta Sci. Math. (Szeged), 19 (1958) 172-187. This work was continued by him in *The rotation of eigenvectors by a perturbation*, J. Math. Anal. Appl., 6 (1963) 159-173, and eventually led to the famous paper by C. Davis and W.M. Kahan, *The rotation of eigenvectors by a perturbation III*, SIAM J. Numer. Anal. 7(1970) 1-46. The CS decomposition in its explicit form as given in Theorem VII.1.6 is due to G.W. Stewart, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Rev., 19(1977) 634-662. It occurs implicitly in the paper by Davis and Kahan cited above as well as in another important paper, A. Björck and G.H. Golub, *Numerical methods for computing angles between linear subspaces*, Math. Comp. 27(1973) 579-594. Further references may be found in an excellent survey article, C.C. Paige and M. Wei, *History and generality of the CS Decomposition.* Linear Algebra Appl., 208/209 (1994) 303-326. Many of the proofs in Section VII.1 are slight modifications of those given by Stewart and Sun in *Matrix Perturbation Theory.*

The importance of the Sylvester equation $AX - XB = Y$ in connection with perturbation of eigenvectors was recognised and emphasized by Davis and Kahan, and in the subsequent paper, G.W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973) 727-764. Almost all results we have derived for this equation are true also in infinite-dimensional Hilbert spaces. More on this equation and its applications, and an extensive bibliography, may be found in R. Bhatia and P. Rosenthal, *How and why to solve the equation* $AX - XB = Y$, Bull. London Math. Soc., 29(1997) to appear.

Theorem VII.2.1 was proved by J.H. Sylvester, *Sur ℓ'equation*

en matrices $px = xq$, C.R. Acad. Sci. Paris 99(1884) 67-71 and 115-116. The proof given here is due to G. Lumer and M. Rosenblum, *Linear operator equations*, Proc. Amer. Math. Soc., 10 (1959) 32-41. The solution (VII.18) is due to E. Heinz, *Beiträge zur Störungstheorie der Spektralzerlegung*, Math. Ann., 123 (1951) 415-438. The general solution (VII.19) is due to M. Rosenblum, *On the operator equation $BX - XA = Q$*, Duke Math. J., 23 (1956) 263-270. Much of the rest of Section VII.2 is based on the paper by R. Bhatia, C. Davis and A. McIntosh, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52/53 (1983) 45-67.

For Hermitian operators, Theorem VII.3.1 was proved in the Davis-Kahan paper cited above. This is very well known among numerical analysts as the $\sin\theta$ theorem and has been used frequently by them. This paper also contains a $\tan\theta$ theorem (see Problem VII.6.6) as well as $\sin 2\theta$ and $\tan 2\theta$ theorems, all for Hermitian operators. Theorem VII.3.4 (for Hermitian operators) is also proved there. The rest of Section VII.3 is based on the paper by Bhatia, Davis, and McIntosh cited above, as is Section VII.4.

Theorem VII.5.1 is due to R.-C.Li, *New perturbation bounds for the unitary polar factor*, SIAM J. Matrix Anal. Appl., 16 (1995) 327-332. Lemmas VII.5.4, VII.5.5 and Corollary VII.5.6 were proved in F. Kittaneh, *On Lipschitz functions of normal operators*, Proc. Amer. Math. Soc., 94 (1985) 416-418. Theorem VII.5.7 is also due to F. Kittaneh, *Inequalities for the Schatten p-norm IV*, Commun. Math. Phys., 106 (1986) 581-585. The inequality (VII.39) was proved earlier by H. Araki and S. Yamagami, *An inequality for the Hilbert-Schmidt norm*, Commun. Math. Phys., 81 (1981) 89-98. Theorem VII.5.9 was proved in R. Bhatia and F. Kittaneh, *On some perturbation inequalities for operators*, Linear Algebra Appl., 106 (1988) 271-279. See also P.A. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT, 12(1972) 99-111.

The beautiful result $c_1 = \frac{\pi}{2}$, and its proof in the Appendix, are taken from B. Sz.-Nagy, *Über die Ungleichung von H. Bohr*, Math. Nachr., 9(1953) 255-259. Problems such as this, are called *minimal extrapolation problems* for the Fourier transform. See H.S. Shapiro, *Topics in Approximation Theory*, Springer Lecture Notes in Mathematics Vol. 187 (1971).

In Theorem VII.2.5, all that is required of f is that $\hat{f}(s) = \frac{1}{s}$ when $s \in \sigma(A) - \sigma(B)$. This fixes the value of $\hat{f}$ only at n^2 points if we are dealing with $n \times n$ matrices. For each n, let $b(n)$ be the smallest constant, for which we have

$$\|X\| \leq \frac{b(n)}{\delta} \|AX - XB\|,$$

whenever A, B are $n \times n$ Hermitian matrices such that dist $(\sigma(A), \sigma(B)) = \delta$. R. McEachin has shown that

$$b(2) = \frac{\sqrt{6}}{2} \approx 1.22474 \text{ (see Example VII.2.10)}$$

$$b(3) = \frac{8+5\sqrt{10}}{18} \approx 1.32285$$

and that

$$b = \lim_{n\to\infty} b(n) = \frac{\pi}{2}.$$

Thus the inequality (VII.26) is sharp with $c_1 = \frac{\pi}{2}$. See, R. McEachin, *A sharp estimate in an operator inequality*, Proc. Amer. Math. Soc., 115 (1992) 161-165 and *Analyzing specific cases of an operator inequality*, Linear Algebra Appl., 208/209 (1994) 343-365.

The quantity $\rho(\mathcal{E},\mathcal{F})$ defined in Problem VII.6.1 is sometimes called the **gap** between $\mathcal{E}$ and $\mathcal{F}$. This and related measures of the distance between two subspaces of a Banach space are used extensively by T. Kato, *Perturbation Theory for Linear Operators*, Chapter 4.

VIII

Spectral Variation of Nonnormal Matrices

In Chapter 6 we saw that if A and B are both Hermitian or both unitary, then the optimal matching distance $d(\sigma(A), \sigma(B))$ is bounded by $\|A-B\|$. We also saw that for arbitrary normal matrices A, B this need not always be true (Example VI.3.13). However, in this case, we do have a slightly weaker inequality $d(\sigma(A), \sigma(B)) \leq 3\|A-B\|$ (Theorem VII.4.1). If one of the matrices A, B is Hermitian and the other is arbitrary, then we can only have an inequality of the form $d(\sigma(A), \sigma(B)) \leq c(n)\|A-B\|$, where $c(n)$ is a constant that grows like $\log n$ (Problems VI.8.8 and VI.8.9).

A more striking change of behaviour takes place if no restriction is placed on either A or B. Let A be the $n \times n$ nilpotent upper Jordan matrix; i.e., the matrix that has all entries 1 on its first diagonal above the main diagonal and all other entries 0. Let B be the matrix obtained from A by adding an entry ε in the bottom left corner. Then the eigenvalues of B are the nth roots of ε. So $d(\sigma(A), \sigma(B)) = \varepsilon^{1/n}$, whereas $\|A-B\| = \varepsilon$. When ε is small, the quantity $\varepsilon^{1/n}$ is much larger. No inequality like $d(\sigma(A), \sigma(B)) \leq c(n)\|A-B\|$ can be true in this case.

In this chapter we will obtain bounds for $d(\sigma(A), \sigma(B))$, where A, B are arbitrary matrices. These bounds are much weaker than the ones for normal matrices. We will also obtain stronger results for matrices that are not normal but have some other special properties.

VIII.1 General Spectral Variation Bounds

Throughout this section, A and B will be two $n \times n$ matrices with eigenvalues $\alpha_1, \ldots, \alpha_n$, and $\beta_1, \ldots, \beta_n$, respectively. In Section VI.3, we introduced the notation

$$s(\sigma(B),\ \sigma(A)) = \max_j \min_i |\alpha_i - \beta_j|. \tag{VIII.1}$$

A bound for this number is given in the following theorem.

Theorem VIII.1.1 *Let A, B be $n \times n$ matrices. Then*

$$s(\sigma(B),\ \sigma(A)) \leq (\|A\| + \|B\|)^{1-1/n} \|A - B\|^{1/n}. \tag{VIII.2}$$

Proof. Let j be the index for which the maximum in the definition (VIII.1) is attained. Choose an orthonormal basis $e_1, \ldots, e_n$ such that $Be_1 = \beta_j e_1$. Then

$$\begin{aligned} [s(\sigma(B),\ \sigma(A))]^n &= [\min_i |\alpha_i - \beta_j|]^n \\ &\leq \prod_{i=1}^{n} |\alpha_i - \beta_j| = |\det(A - \beta_j I)| \\ &\leq \|(A - \beta_j I)e_1\| \cdots \|(A - \beta_j I)e_n\| \end{aligned}$$

by Hadamard's inequality (Exercise I.1.3). The first factor on the right-hand side of the above inequality can be written as $\|(A - B)e_1\|$ and is, therefore, bounded by $\|A - B\|$. The remaining $n-1$ factors can be bounded as

$$\|(A - \beta_j I)e_k\| \leq \|Ae_k\| + |\beta_j| \leq \|A\| + \|B\|, \text{ for } k = 2, 3, \ldots, n.$$

This is adequate to derive (VIII.2). ∎

Example VIII.1.2 *Let $A = -B = I$. Then the two sides of (VIII.2) are equal.*

Compare this theorem with Theorem VI.3.3.

Since the right-hand side of (VIII.2) is symmetric in A and B, we have a bound for the Hausdorff distance as well:

$$h(\sigma(A),\ \sigma(B)) \leq (\|A\| + \|B\|)^{1-1/n} \|A - B\|^{1/n}. \tag{VIII.3}$$

Exercise VIII.1.3 *A bound for the optimal matching distance $d(\sigma(A),\ \sigma(B))$ can be derived from Theorem VIII.1.1. The argument is similar to the one used in Problem VI.8.6 and is outlined below.*

(i) Fix A, and for any B let

$$\varepsilon(B) = (2M)^{1-1/n} \|A - B\|^{1/n},$$

where $M = \max(\|A\|, \|B\|)$. Let $\alpha_1, \ldots, \alpha_n$ be the eigenvalues of A. Let $\bar{D}(\alpha_i, \varepsilon(B))$ be the closed disk with radius $\varepsilon(B)$ and centre α_i. Then, Theorem VIII.1.1 says that $\sigma(B)$ is contained in the set D obtained by taking the union of these disks.

(ii) Let $A(t) = (1-t)A+tB$, $0 \le t \le 1$. Then $A(0) = A$, $A(1) = B$, and $\varepsilon(A(t)) \le \varepsilon(B)$ for all t. Thus, for each $0 \le t \le 1$, $\sigma(A(t))$ is contained in D.

(iii) Since the n eigenvalues of $A(t)$ are continuous functions of t, each connected component of D contains as many eigenvalues of B as of A.

(iv) Use the Matching Theorem to show that this implies

$$d(\sigma(A), \sigma(B)) \le (2n-1)(2M)^{1-1/n}\|A-B\|^{1/n}. \tag{VIII.4}$$

(v) Interchange the roles of A and B and use the result of Problem II.5.10 to obtain the stronger inequality

$$d(\sigma(A), \sigma(B)) \le n(2M)^{1-1/n}\|A-B\|^{1/n}. \tag{VIII.5}$$

The example given in the introduction shows that the exponent $1/n$ occurring on the right-hand side of (VIII.5) is necessary. But then homogeneity considerations require the insertion of another factor like $(2M)^{1-1/n}$. However, the first factor n on the right-hand side of (VIII.5) can be replaced by a much smaller constant factor. This is shown in the next theorem. We will use a classical result of Chebyshev used frequently in approximation theory: if p is any monic polynomial of degree n, then

$$\max_{0\le t\le 1} |p(t)| \ge \frac{1}{2^{2n-1}}. \tag{VIII.6}$$

(This can be found in standard texts such as P. Henrici, *Elements of Numerical Analysis*, Wiley, 1964, p. 194; T.J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, 1981, p. 31.) The following lemma is a generalisation of this inequality.

Lemma VIII.1.4 *Let Γ be a continuous curve in the complex plane with endpoints a and b. If p is any monic polynomial of degree n, then*

$$\max_{\lambda\in\Gamma} |p(\lambda)| \ge \frac{|b-a|^n}{2^{2n-1}}. \tag{VIII.7}$$

Proof. Let L be the straight line through a and b and S the segment of L between a and b:

$$\begin{aligned} L &= \{z : z = a + t(b-a), \ \ t \in \mathbb{R}\} \\ S &= \{z : z = a + t(b-a), \ \ 0 \le t \le 1\}. \end{aligned}$$

For every point z in $\mathbb{C}$, let z' denote its orthogonal projection onto L. Then $|z-w| \ge |z'-w'|$ for all z and w.

Let λ_i, $i = 1, \ldots, n$, be the roots of p. Let $\lambda_i' = a + t_i(b-a)$, where $t_i \in \mathbb{R}$, and let $z = a + t(b-a)$ be any point on L. Then

$$\prod_{i=1}^{n} |z - \lambda_i'| = \prod_{i=1}^{n} |(t-t_i)(b-a)| = |b-a|^n \prod_{i=1}^{n} |t - t_i|.$$

From this and the inequality (VIII.6) applied to the polynomial $\prod_{i=1}^{n}(t-t_i)$, we can conclude that there exists a point z_0 on S for which

$$\prod_{i=1}^{n} |z_0 - \lambda_i'| \geq \frac{|b-a|^n}{2^{2n-1}}.$$

Since Γ is a continuous curve joining a and b, $z_0 = \lambda_0'$ for some $\lambda_0 \in \Gamma$. Since $|\lambda_0 - \lambda_i| \geq |\lambda_0' - \lambda_i'|$, we have shown that there exists a point λ_0 on Γ such that $|p(\lambda_0)| = \prod_{i=1}^{n} |\lambda_0 - \lambda_i| \geq \frac{|b-a|^n}{2^{2n-1}}$. ∎

Theorem VIII.1.5 *Let A and B be two $n \times n$ matrices. Then*

$$d(\sigma(A), \sigma(B)) \leq 4(\|A\| + \|B\|)^{1-1/n} \|A - B\|^{1/n}. \tag{VIII.8}$$

Proof. Let $A(t) = (1-t)A + tB$, $0 \leq t \leq 1$. The eigenvalues of $A(t)$ trace n continuous curves in the plane as t changes from 0 to 1. The initial points of these curves are the eigenvalues of A, and their final points are those of B. So, to prove (VIII.8) it suffices to show that if Γ is one of these curves and a and b are the endpoints of Γ, then $|a-b|$ is bounded by the right-hand side of (VIII.8).

Assume that $\|A\| \leq \|B\|$ without any loss of generality. By Lemma VIII.1.4, there exists a point λ_0 on Γ such that

$$|\det(A - \lambda_0 I)| \geq \frac{|b-a|^n}{2^{2n-1}}.$$

Choose $0 \leq t_0 \leq 1$ such that λ_0 is an eigenvalue of $(1-t_0)A + t_0 B$. In the proof of Theorem VIII.1.1 we have seen that if X, Y are any two $n \times n$ matrices and if λ is an eigenvalue of Y, then

$$|\det(X - \lambda I)| \leq \|X - Y\| (\|X\| + \|Y\|)^{n-1}.$$

Choose $X = A$ and $Y = (1-t_0)A + t_0 B$. This gives $\frac{|b-a|^n}{2^{2n-1}} \leq |\det(A - \lambda_0 I)| \leq \|A - B\|(\|A\| + \|B\|)^{n-1}$. Taking nth roots, we obtain the desired conclusion. ∎

Note that we have, in fact, shown that the factor 4 in the inequality (VIII.8) can be replaced by the smaller number $4 \times 2^{-1/n}$. A further improvement is possible; see the Notes at the end of the chapter. However, the best possible inequality of this type is not yet known.

VIII.2 Perturbation of Roots of Polynomials

The ideas used above also lead to bounds for the distance between the roots of two polynomials. This is discussed below.

Lemma VIII.2.1 *Let $f(z) = z^n + a_1 z^{n-1} + \cdots + a_n$ be any monic polynomial. Let*

$$\mu = 2 \max_{1\le k\le n} |a_k|^{1/k}. \tag{VIII.9}$$

Then all the roots of f are bounded (in absolute value) by μ.

Proof. If $|z| > \mu$, then

$$\begin{aligned} \left|\frac{f(z)}{z^n}\right| &= \left|1 + \frac{a_1}{z} + \cdots + \frac{a_n}{z^n}\right| \\ &\ge 1 - \left|\frac{a_1}{z}\right| - \left|\frac{a_2}{z^2}\right| - \cdots - \left|\frac{a_n}{z^n}\right| \\ &> 1 - \frac{1}{2} - \frac{1}{2^2} - \cdots - \frac{1}{2^n} \\ &> 0. \end{aligned}$$

Such z cannot, therefore, be a root of f. ■

Let $\alpha_1, \ldots, \alpha_n$ be the roots of a monic polynomial f. We will denote by Root f the unordered n-tuple $\{\alpha_1, \ldots, \alpha_n\}$ as well as the subset of the plane whose elements are the roots of f. We wish to find bounds for the optimal matching distance $d(\text{Root } f, \text{Root } g)$ in terms of the distance between the coefficients of two monic polynomials f and g. Let

$$\begin{aligned} f(z) &= z^n + a_1 z^{n-1} + \cdots + a_n, \\ g(z) &= z^n + b_1 z^{n-1} + \cdots + b_n \end{aligned} \tag{VIII.10}$$

be two polynomials. Let

$$\gamma = 2 \max_{1\le k\le n} \max(|a_k|^{1/k}, |b_k|^{1/k}), \tag{VIII.11}$$

$$\Theta(f,g) = \left\{ \sum_{k=1}^{n} |a_k - b_k| \gamma^{n-k} \right\}^{1/n}. \tag{VIII.12}$$

The bounds given below are in terms of these quantities.

Theorem VIII.2.2 *Let f, g be two monic polynomials as in (VIII.10). Then*

$$s(\text{Root } f, \text{Root } g) \le \sum_{k=1}^{n} |a_k - b_k| \mu^{n-k}, \tag{VIII.13}$$

where μ is given by (VIII.9).

Proof. We have

$$f(z) - g(z) = \sum_{k=1}^{n} (a_k - b_k) z^{n-k}.$$

So, if α is any root of f, then, by Lemma VIII.2.1,

$$\begin{aligned} |g(\alpha)| &\leq \sum_{k=1}^{n} |a_k - b_k| \ |\alpha|^{n-k} \\ &\leq \sum_{k=1}^{n} |a_k - b_k| \ \mu^{n-k}. \end{aligned}$$

If the roots of g are $\beta_1, \ldots, \beta_n$, this says that

$$\prod_{j=1}^{n} |\alpha - \beta_j| \leq \sum_{k=1}^{n} |a_k - b_k| \mu^{n-k}.$$

So,

$$\min_j |\alpha - \beta_j| \leq \left\{ \sum_{k=1}^{n} |a_k - b_k| \mu^{n-k} \right\}^{1/n}.$$

This proves the theorem. ■

Corollary VIII.2.3 *The Hausdorff distance between the roots of f and g is bounded as*

$$h(\text{Root} f, \text{Root} g) \leq \Theta(f, g). \tag{VIII.14}$$

Theorem VIII.2.4 *The optimal matching distance between the roots of f and g is bounded as*

$$d(\text{Root}\, f, \text{Root}\, g) \leq 4\, \Theta(f, g). \tag{VIII.15}$$

Proof. The argument is similar to that used in proving Theorem VIII.1.5. Let $f_t = (1-t)f + tg, 0 \leq t \leq 1$. If λ is a root of f_t, then by Lemma VIII.2.1, $|\lambda| \leq \gamma$ and we have

$$\begin{aligned} |f(\lambda)| &= |t(f(\lambda) - g(\lambda))| \leq |f(\lambda) - g(\lambda)| \\ &\leq \sum_{k=1}^{n} |a_k - b_k| \ |\lambda|^{n-k} \leq [\Theta(f, g)]^n. \end{aligned}$$

The roots of f_t trace n continuous curves as t changes from 0 to 1. The initial points of these curves are the roots of f, and the final points are the roots of g. Let Γ be any one of these curves, and let a, b be its endpoints. Then, by Lemma VIII.1.4, there exists a point λ on Γ such that

$$|f(\lambda)| \geq \frac{|a - b|^n}{2^{2n-1}}.$$

This shows that

$$|a-b| \leq 4 \times 2^{-1/n}\, \Theta(f,g),$$

and that is enough for proving the theorem. ∎

Exercise VIII.2.5 *Let $h=(f+g)/2$. Then any convex combination of f and g can also be expressed as $h+t(f-g)$ for some t with $|t| \leq \frac{1}{2}$. Use this to show that*

$$d(\text{Root}\, f, \text{Root}\, g) \leq 4^{1-1/n}\, \Theta(f,g). \tag{VIII.16}$$

The first factor in the above inequality can be reduced further; see the Notes at the end of the chapter. However, it is not known what the optimal value of this factor is. It is known that no constant smaller than 2 can replace this factor if the inequality is to be valid for all degrees n.

Note that the only property of γ used in the proof is that it is an upper bound for all the roots of the polynomial $(1-t)f+tg$. Any other constant with this property could be used instead.

Exercise VIII.2.6 *In Problem I.6.11, a bound for the distance between the coefficients of the characteristic polynomials of two matrices was obtained. Use that and the combinatorial identity*

$$\sum_{k=0}^{n} k\binom{n}{k} = n2^{n-1}$$

to show that for any two $n \times n$ matrices A, B

$$d(\sigma(A),\, \sigma(B)) \leq n^{1/n}(8M)^{1-1/n}\|A-B\|^{1/n}, \tag{VIII.17}$$

where $M = \max(\|A\|,\ \|B\|)$. This is weaker than the bound obtained in Theorem VIII.1.5.

VIII.3 Diagonalisable Matrices

A matrix A is said to be **diagonalisable** if it is similar to a diagonal matrix; i.e., if there exists an invertible matrix S and a diagonal matrix D such that $A = SDS^{-1}$. This is equivalent to saying that there are n linearly independent vectors in $\mathbb{C}^n$ that are eigenvectors for A. If S is unitary (or the eigenvectors of A orthonormal), A is normal. In this section we will derive some perturbation bounds for diagonalisable matrices. These are natural generalisations of some results obtained for normal matrices in Chapter 6.

The **condition number** of an invertible matrix S is defined as

$$\text{cond}(S) = \|S\|\ \|S^{-1}\|.$$

Note that $\text{cond}(S) \geq 1$, and $\text{cond}(S) = 1$ if and only if S is a scalar multiple of a unitary matrix.

Our first theorem is a generalisation of Theorem VI.3.3.

Theorem VIII.3.1 *Let $A = SDS^{-1}$, where D is a diagonal matrix and S an invertible matrix. Then, for any matrix B,*

$$s(\sigma(B), \sigma(A)) \leq \text{cond}(S)\|A - B\|. \qquad \text{(VIII.18)}$$

Proof. The proof of Theorem VI.3.3 can be modified to give a proof of this. Let $\varepsilon = \text{cond}(S)\|A - B\|$. We want to show that if β is any eigenvalue of B, then β is within a distance ε of some eigenvalue α_j of A. By applying a translation, we may assume that $\beta = 0$. If none of the α_j is within a distance ε of this, then A is invertible and

$$\begin{aligned} \|A^{-1}\| = \|SD^{-1}S^{-1}\| &\leq \text{cond}(S)\|D^{-1}\| < \frac{\text{cond}(S)}{\varepsilon} \\ &= \frac{1}{\|A - B\|}. \end{aligned}$$

So,

$$\|A^{-1}(B - A)\| \leq \|A^{-1}\| \, \|B - A\| < 1.$$

Hence, $I + A^{-1}(B - A)$ is invertible, and so is $B = A(I + A^{-1}(B - A))$. But then B could not have had a zero eigenvalue. ∎

Note that the properties of the operator norm used above are (i) $I + A$ is invertible if $\|A\| < 1$; (ii) $\|AB\| \leq \|A\| \, \|B\|$ for all A, B; (iii) $\|D\| = \max |d_i|$ if $D = \text{diag}(d_1, \ldots, d_n)$. There are several other norms that satisfy these properties. For example, norms induced by the p-norms on $\mathbb{C}^n$, $1 \leq p \leq \infty$, all have these three properties. So, the inequality (VIII.18) is true for a large class of norms.

Exercise VIII.3.2 *Using continuity arguments and the Matching Theorem show that if A and B are as in Theorem VIII.3.1, then*

$$d(\sigma(A), \sigma(B)) \leq (2n - 1)\text{cond}(S)\|A - B\|.$$

If B is also diagonalisable and $B = TD'T^{-1}$, then

$$d(\sigma(A), \sigma(B)) \leq n \, \text{cond}(S)\text{cond}(T)\|A - B\|.$$

An inequality stronger than this will be proved below by other means.

Theorem VIII.3.1 also follows from the following theorem. Both of these are called the **Bauer-Fike Theorems**.

Theorem VIII.3.3 *Let S be an invertible matrix. If β is an eigenvalue of B but not of A, then*

$$\|S(A - \beta I)^{-1}S^{-1}\|^{-1} \leq \|S(A - B)S^{-1}\|. \qquad \text{(VIII.19)}$$

Proof. We can write

$$\begin{aligned} S(B-\beta I)S^{-1} &= S[(A-\beta I)+B-A]S^{-1} \\ &= S(A-\beta I)S^{-1}\{I+S(A-\beta I)^{-1}S^{-1}\cdot S(B-A)S^{-1}\}. \end{aligned}$$

Note that the matrix on the left-hand side of this equation is not invertible. Since $A-\beta I$ is invertible, the matrix inside the braces on the right-hand side is not invertible. Hence,

$$\begin{aligned} 1 &\le \|S(A-\beta I)^{-1}S^{-1}\cdot S(B-A)S^{-1}\| \\ &\le \|S(A-\beta I)^{-1}S^{-1}\|\ \|S(B-A)S^{-1}\|. \end{aligned}$$

This proves the theorem. ■

We now obtain, for diagonalisable matrices, analogues of some of the major perturbation bounds derived in earlier chapters for Hermitian matrices and for normal matrices. Some auxiliary theorems about norms of commutators are proved first.

Theorem VIII.3.4 *Let A, B be Hermitian operators, and let Γ be a positive operator whose smallest eigenvalue is γ (i.e., $\Gamma \ge \gamma I \ge 0$). Then*

$$|||A\Gamma-\Gamma B||| \ge \gamma |||A-B||| \qquad \text{(VIII.20)}$$

for every unitarily invariant norm.

Proof. Let $T = A\Gamma - \Gamma B$, and let $Y = T + T^*$. Then

$$Y = (A-B)\Gamma + \Gamma(A-B).$$

This is the Sylvester equation that we studied in Chapter 7. From Theorem VII.2.12, we get

$$2\gamma|||A-B||| \le |||Y||| \le 2|||T||| = 2|||A\Gamma-\Gamma B|||.$$

This proves the theorem. ■

Corollary VIII.3.5 *Let A, B be any two operators, and let Γ be a positive operator, $\Gamma \ge \gamma I \ge 0$. Then*

$$|||(A\Gamma-\Gamma B)\oplus(A^*\Gamma-\Gamma B^*)||| \ge \gamma|||(A-B)\oplus(A-B)||| \qquad \text{(VIII.21)}$$

for every unitarily invariant norm.

Proof. This follows from (VIII.20) applied to the Hermitian operators $\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix}$, and the positive operator $\begin{pmatrix} \Gamma & 0 \\ 0 & \Gamma \end{pmatrix}$. ■

Corollary VIII.3.6 *Let A and B be unitary operators, and let Γ be a positive operator, $\Gamma \geq \gamma I \geq 0$. Then, for every unitarily invariant norm,*

$$|||A\Gamma - \Gamma B||| \geq \gamma |||A - B|||. \tag{VIII.22}$$

Proof. If A and B are unitary, then

$$s_j(A\Gamma - \Gamma B) = s_j(\Gamma B^* - A^*\Gamma) = s_j(A^*\Gamma - \Gamma B^*).$$

Thus the operator $(A\Gamma - \Gamma B) \oplus (A^*\Gamma - \Gamma B^*)$ has the same singular values as those of $A\Gamma - \Gamma B$, each counted twice. From this we see that (VIII.21) implies (VIII.22) for all Ky Fan norms, and hence for all unitarily invariant norms. ∎

Corollary VIII.3.7 *Let A, B be normal operators, and let Γ be a positive operator, $\Gamma \geq \gamma I \geq 0$. Then*

$$\|A\Gamma - \Gamma B\|_2 \geq \gamma \|A - B\|_2. \tag{VIII.23}$$

Proof. Suppose that A is normal and its eigenvalues are $\alpha_1, \ldots, \alpha_n$. Then (choosing an orthonormal basis in which A is diagonal) one sees that for every X

$$\|AX - XA\|_2^2 = \sum_{i,j} |\alpha_i - \alpha_j|^2 |x_{ij}|^2 = \|A^*X - XA^*\|_2^2.$$

If A, B are normal, then, applying this to $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ in place of A and $\begin{pmatrix} 0 & X \\ 0 & 0 \end{pmatrix}$ in place of X, one obtains

$$\|AX - XB\|_2 = \|A^*X - XB^*\|_2. \tag{VIII.24}$$

Using this, the inequality (VIII.23) can be derived from (VIII.21). ∎

A famous theorem (called the **Fuglede-Putnam Theorem**, valid in Hilbert spaces of finite or infinite dimensions) says that if A and B are normal, then for any operator $X, AX = XB$ if and only if $A^*X = XB^*$. The equality (VIII.24) says much more than this.

Example VIII.3.8 *For normal A, B, the inequality (VIII.23) is not always true if the Hilbert-Schmidt norm is replaced by the operator norm. A numerical example illustrating this is given below. Let*

$$\Gamma = \mathrm{diag}(0.6384,\ 0.6384,\ 1.0000),$$

$$A = \begin{pmatrix} -0.5205 - 0.1642i & 0.1042 - 0.3618i & -0.1326 - 0.0260i \\ -0.1299 + 0.1709i & 0.4218 + 0.4685i & -0.5692 - 0.3178i \\ 0.2850 - 0.1808i & -0.3850 - 0.4257i & -0.2973 - 0.1715i \end{pmatrix},$$

$$B = \begin{pmatrix} -0.6040+0.1760i & 0.5128-0.2865i & 0.1306+0.0154i \\ 0.0582+0.2850i & 0.0154+0.4497i & -0.5001-0.2833i \\ 0.4081-0.3333i & -0.0721-0.2545i & -0.2686+0.0247i \end{pmatrix}.$$

Then

$$\frac{\|A\Gamma - \Gamma B\|}{\gamma\|A-B\|} = 0.8763.$$

Theorem VIII.3.4 and Corollary VIII.3.7 are used in the proofs below. Alternate proofs of both these results are sketched in the problems. These proofs do not draw on the results in Chapter 7.

Theorem VIII.3.9 *Let A, B be any two matrices such that $A = SD_1S^{-1}$, $B = TD_2T^{-1}$, where S, T are invertible matrices and D_1, D_2 are real diagonal matrices. Then*

$$|||\mathrm{Eig}^{\downarrow}(A) - \mathrm{Eig}^{\downarrow}(B)||| \leq [\mathrm{cond}(S)\mathrm{cond}(T)]^{1/2}|||A-B||| \qquad \text{(VIII.25)}$$

for every unitarily invariant norm.

Proof. When A, B are Hermitian, this has already been proved; see (IV.62). This special case will be used to prove the general result.

We can write

$$A - B = SD_1S^{-1} - TD_2T^{-1} = S(D_1S^{-1}T - S^{-1}TD_2)T^{-1}.$$

Hence,

$$|||D_1S^{-1}T - S^{-1}TD_2||| = |||S^{-1}(A-B)T||| \leq \|S^{-1}\|\ |||A-B|||\ \|T\|.$$

We could also write

$$A - B = T(T^{-1}SD_1 - D_2T^{-1}S)S^{-1}$$

and get

$$|||T^{-1}SD_1 - D_2T^{-1}S||| \leq \|T^{-1}\|\ |||A-B|||\ \|S\|.$$

Let $S^{-1}T$ have the singular value decomposition $S^{-1}T = U\Gamma V$. Then

$$\begin{aligned} |||D_1S^{-1}T - S^{-1}TD_2||| &= |||D_1U\Gamma V - U\Gamma VD_2||| \\ &= |||U^*D_1U\Gamma - \Gamma VD_2V^*||| = |||A'\Gamma - \Gamma B'|||, \end{aligned}$$

where $A' = U^*D_1U$ and $B' = VD_2V^*$ are Hermitian matrices. Note that $T^{-1}S = V^*\Gamma^{-1}U^*$. So, by the same argument,

$$|||T^{-1}SD_1 - D_2T^{-1}S||| = |||\Gamma^{-1}A' - B'\Gamma^{-1}|||.$$

We have, thus, two inequalities

$$\alpha|||A-B||| \geq |||A'\Gamma - \Gamma B'|||,$$

and

$$\beta|||A - B||| \geq |||A'\Gamma^{-1} - \Gamma^{-1}B'|||,$$

where $\alpha = \|S^{-1}\|\,\|T\|$, $\beta = \|T^{-1}\|\,\|S\|$. Combining these two inequalities and using the triangle inequality, we have

$$2|||A - B||| \geq |||A'\left(\frac{\Gamma}{\alpha} + \frac{\Gamma^{-1}}{\beta}\right) - \left(\frac{\Gamma}{\alpha} + \frac{\Gamma^{-1}}{\beta}\right)B'|||.$$

The operator inequality $\left[\left(\frac{\Gamma}{\alpha}\right)^{1/2} - \left(\frac{\Gamma^{-1}}{\beta}\right)^{1/2}\right]^2 \geq 0$ implies that $\frac{\Gamma}{\alpha} + \frac{\Gamma^{-1}}{\beta} \geq \frac{2}{(\alpha\beta)^{1/2}}I$. Hence, by Theorem VIII.3.4,

$$2|||A - B||| \geq \frac{2}{(\alpha\beta)^{1/2}}|||A' - B'|||.$$

But A' and B' are Hermitian matrices with the same eigenvalues as those of A and B, respectively. Hence, by the result for Hermitian matrices that was mentioned at the beginning,

$$|||A' - B'||| \geq |||\mathrm{Eig}^{\downarrow}(A) - \mathrm{Eig}^{\downarrow}(B)|||.$$

Combining the three inequalities above leads to (VIII.25). ■

Theorem VIII.3.10 *Let A, B be any two matrices such that $A = SD_1S^{-1}$, $B = TD_2T^{-1}$, where S, T are invertible matrices and D_1, D_2 are diagonal matrices. Then*

$$d_2(\sigma(A),\ \sigma(B)) \leq [\mathrm{cond}(S)\mathrm{cond}(T)]^{1/2}\|A - B\|_2. \qquad \text{(VIII.26)}$$

Proof. When A, B are normal, this is just the Hoffman-Wielandt inequality; see (VI.34). The general case can be obtained from this using the inequality (VIII.23). The argument is the same as in the proof of the preceding theorem. ■

Theorems VIII.3.9 and VIII.3.10 do reduce to the ones proved earlier for Hermitian and normal matrices. However, neither of them gives tight bounds. Even in the favourable case when A and B commute, the left-hand side of (VIII.24) is generally smaller than $|||A - B|||$, and this is aggravated further by introducing the condition number coefficients.

Exercise VIII.3.11 *Let A and B be as in Theorem VIII.3.10. Suppose that all eigenvalues of A and B have modulus 1. Show that*

$$d_{|||\cdot|||}(\sigma(A), \sigma(B)) \leq \frac{\pi}{2}[\mathrm{cond}(S)\mathrm{cond}(T)]^{1/2}|||A - B||| \qquad \text{(VIII.27)}$$

for all unitarily invariant norms. For the special case of the operator norm, the factor $\frac{\pi}{2}$ above can be replaced by 1.

[Hint: Use Corollary VIII.3.6 and the theorems on unitary matrices in Chapter 6.]

VIII.4 Matrices with Real Eigenvalues

In this section we will consider a collection $\mathcal{R}$ of matrices that has two special properties: $\mathcal{R}$ is a real vector space and every element of $\mathcal{R}$ has only real eigenvalues. The set of all Hermitian matrices is an example of such a collection. Another example is given below. Such families of matrices arise in the study of vectorial hyperbolic differential equations. The behaviour of the eigenvalues of such a family has some similarities to that of Hermitian matrices. This is studied below.

Example VIII.4.1 *Fix a block decomposition of matrices in which all diagonal blocks are square. Let $\mathcal{R}$ be the set of all matrices that are block upper triangular in this decomposition and whose diagonal blocks are Hermitian. Then $\mathcal{R}$ is a real vector space (of real dimension n^2) and every element of $\mathcal{R}$ has real eigenvalues.*

In this book we have called a matrix positive if it is Hermitian and all its eigenvalues are nonnegative. A matrix A will be called **laxly positive** if all eigenvalues of A are nonnegative. This will be written symbolically as $0 \leq^L A$. If all eigenvalues of A are positive, we will say A is **strictly laxly positive**. We say $A \leq^L B$ if $B - A$ is laxly positive.

We will see below that if $\mathcal{R}$ is a real vector space of matrices each of which has only real eigenvalues, then the laxly positive elements form a convex cone in $\mathcal{R}$. So, the order $\leq^L$ defines a partial order on $\mathcal{R}$.

Given two matrices A and B, we say that λ is an **eigenvalue of A with respect to B** if there exists a nonzero vector x such that $Ax = \lambda Bx$. Thus, eigenvalues of A with respect to B are the n roots of the equation $\det(A - \lambda B) = 0$. These are also called **generalised eigenvalues**.

Lemma VIII.4.2 *Let A, B be two matrices such that every real linear combination of A and B has real eigenvalues. Suppose B is strictly laxly positive. Then for every real λ, $-A + \lambda I$ has real eigenvalues with respect to B.*

Proof. We have to show that for any real λ the equation

$$\det(-A + \lambda I - \mu B) = 0 \tag{VIII.28}$$

is satisfied by n real μ.

Let μ be any given real number. Then, by hypothesis, there exist n real λ that satisfy (VIII.28), namely the eigenvalues of $A + \mu B$. Denote these λ as $\varphi_j(\mu)$ and arrange them so that $\varphi_1(\mu) \geq \varphi_2(\mu) \geq \cdots \geq \varphi_n(\mu)$. We have

$$\det(-A + \lambda I - \mu B) = \prod_{k=1}^{n} (\lambda - \varphi_k(\mu)). \tag{VIII.29}$$

By the results of Section VI.1, each $\varphi_k(\mu)$ is continuous as a function of μ. For large μ, $\frac{1}{\mu}(A + \mu B)$ is close to B. So, $\frac{1}{\mu}\varphi_k(\mu)$ approaches $\lambda_k^{\downarrow}(B)$ as

$\mu \to \infty$, and $\lambda_k^{\uparrow}(B)$ as $\mu \to -\infty$. Since B is strictly laxly positive, this implies that $\varphi_k(\mu) \to \pm\infty$ as $\mu \to \pm\infty$.

So, every λ in $\mathbb{R}$ is in the range of φ_k for each $k = 1, 2, \ldots, n$. Thus, for each λ, there exist n real μ that satisfy (VIII.28). ∎

Proposition VIII.4.3 *Let A, B be two matrices such that every real linear combination of A and B has real eigenvalues. Suppose A is (strictly) laxly negative. Then every eigenvalue of $A + iB$ has (strictly) negative real part.*

Proof. Let $\mu = \mu_1 + i\mu_2$ be an eigenvalue of $A + iB$. Then det $(A + iB - \mu_1 I - i\mu_2 I) = 0$. Multiply this by i^n to get

$$\det[(-B + \mu_2 I) + i(A - \mu_1 I)] = 0.$$

So the matrix $-B + \mu_2 I$ has an eigenvalue $-i$ with respect to the matrix $A - \mu_1 I$, and it has an eigenvalue i with respect to the matrix $-(A - \mu_1 I)$.

By hypothesis, every real linear combination of $A - \mu_1 I$ and B has real eigenvalues. Hence, by Lemma VIII.4.2, $A - \mu_1 I$ cannot be either strictly laxly positive or strictly laxly negative. In other words,

$$\lambda_n^{\downarrow}(A) \le \mu_1 \le \lambda_1^{\downarrow}(A).$$

This proves the proposition. ∎

Exercise VIII.4.4 *With notations as in the above proof, show that*

$$\lambda_n^{\downarrow}(B) \le \mu_2 \le \lambda_1^{\downarrow}(B).$$

Theorem VIII.4.5 *Let $\mathcal{R}$ be a real vector space whose elements are matrices with real eigenvalues. Let $A, B \in \mathcal{R}$ and let $A \le^L B$. Then $\lambda_k^{\downarrow}(A) \le \lambda_k^{\downarrow}(B)$ for $k = 1, 2, \ldots, n$.*

Proof. We will prove a more general statement: if $A, B \in \mathcal{R}$ and $0 \le^L B$, then $\lambda_k^{\downarrow}(A + \mu B)$ is a monotonically increasing function of the real variable μ. It is enough to prove this when $0 <^L B$; the general case follows by continuity. In the notation of Lemma VIII.4.2, $\lambda_k^{\downarrow}(A + \mu B) = \varphi_k(\mu)$. Suppose $\varphi_k(\mu)$ decreases in some interval. Then we can choose a real number λ such that $\lambda - \varphi_k(\mu)$ increases from a negative to a positive value in this interval. Since $\varphi_k(\mu) \to \pm\infty$ as $\mu \to \pm\infty$, for this value of $\lambda, \lambda - \varphi_k(\mu)$ vanishes for at least three values of μ. So, in the representation (VIII.29) this factor contributes at least three zeroes. The remaining factors contribute at least one zero each. So, for this λ, the equation (VIII.28) has at least $n+2$ roots μ. This is impossible. ∎

Theorem VIII.4.6 *Let $\mathcal{R}$ be a real vector space whose elements are matrices with real eigenvalues. Let $A, B \in \mathcal{R}$. Then*

$$\lambda_k^{\downarrow}(A) + \lambda_n^{\downarrow}(B) \le \lambda_k^{\downarrow}(A + B) \le \lambda_k^{\downarrow}(A) + \lambda_1^{\downarrow}(B) \qquad \text{(VIII.30)}$$

for $k = 1, 2, \ldots, n$.

Proof. The matrix $B - \lambda_n^{\downarrow}(B)I$ is laxly positive. So, by the argument in the proof of the preceding theorem, $\lambda_k^{\downarrow}(A + \mu B) - \mu\lambda_n^{\downarrow}(B)$ is a monotonically increasing function of μ. Choose $\mu = 0, 1$ to get the first inequality in (VIII.30). The same argument shows that $\lambda_k^{\downarrow}(A+\mu B) - \mu\lambda_1^{\downarrow}(B)$ is a monotonically decreasing function of μ. This leads to the second inequality. ■

Corollary VIII.4.7 *On the vector space $\mathcal{R}$, the function $\lambda_1^{\downarrow}(A)$ is convex and the function $\lambda_n^{\downarrow}(A)$ is concave in the argument A.*

Theorem VIII.4.8 *Let A and B be two matrices such that all real linear combinations of A and B have real eigenvalues. Then*

$$\max_{1\le k\le n} |\lambda_k^{\downarrow}(A) - \lambda_k^{\downarrow}(B)| \le \operatorname{spr}(A - B) \le \|A - B\|. \tag{VIII.31}$$

Proof. Let $\mathcal{R}$ be the real vector space generated by A and B. By Theorem VIII.4.6,

$$\lambda_k^{\downarrow}(A) + \lambda_n^{\downarrow}(B - A) \le \lambda_k^{\downarrow}(B) \le \lambda_k^{\downarrow}(A) + \lambda_1^{\downarrow}(B - A).$$

So,

$$\begin{aligned} |\lambda_k^{\downarrow}(B) - \lambda_k^{\downarrow}(A)| &\le \max(|\lambda_1^{\downarrow}(B - A)|,\ |\lambda_n^{\downarrow}(B - A)|) \\ &= \operatorname{spr}(A - B) \le \|A - B\|. \end{aligned}$$

■

Note that Weyl's Perturbation Theorem is included in this as a special case.

Exercise VIII.4.9 *Show that if only A, B and $A+B$ are assumed to have real eigenvalues, then the inequality (VIII.31) might not be true.*

VIII.5 Eigenvalues with Symmetries

We have remarked earlier that the exponent $1/n$ occurring in the bound (VIII.8) is unavoidable. However, if A and B are restricted to some special classes, this can be improved. In this section we identify some useful classes of matrices where this exponent can be improved (though not eliminated altogether). These are matrices whose eigenvalues appear as pairs $\pm\lambda$ or, more generally, as tuples $\{\lambda, \omega\lambda, \ldots, \omega^{p-1}\lambda\}$, where ω is a pth root of unity. We will give interesting examples of large classes of such matrices, and then show how this symmetric distribution of their eigenvalues can be exploited to get better bounds.

Example VIII.5.1 *Let A^T denote the transpose of a matrix. A complex matrix is called* **symmetric** *if $A^T = A$ and* **skew-symmetric** *if $A^T = -A$. If A is a skew-symmetric matrix, then λ is an eigenvalue of A if and only if $-\lambda$ is. The class of all such matrices forms a Lie algebra. This is the Lie algebra associated with the complex orthogonal group.*

Example VIII.5.2 *If A^T is similar to $-A$, then, clearly, λ is an eigenvalue of A if and only if $-\lambda$ is. The Lie algebra corresponding to the symplectic Lie group contains matrices that have this property. Let n be an even number $n = 2r$. Let $J = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$, where I is the identity matrix of order r. Let A be an $n \times n$ matrix such that $A^T = -JAJ^{-1}$. It is easy to see that we can then write*

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & -A_1^T \end{pmatrix}$$

where A_1, A_2, A_3 are $r \times r$ matrices of which A_2 and A_3 are skew-symmetric. The collection of all such matrices is the Lie algebra associated with the symplectic group.

Example VIII.5.3 *Let X be a matrix of order $n = pr$ having a special form*

$$X = \begin{pmatrix} 0 & A_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & A_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & & \cdots \\ \cdots & \cdots & \cdots & \cdots & & \cdots \\ 0 & 0 & 0 & 0 & \cdots & A_{p-1} \\ A_p & 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

where $A_1, \ldots, A_p$ are matrices of order r. Let $Y = \mathrm{diag}(I_r, \omega I_r, \ldots, \omega^{p-1} I_r)$, where ω is the primitive pth root of unity. Then $Y^{-1}XY = \omega X$. So, if λ is an eigenvalue of X, then so are $\omega\lambda, \omega^2\lambda, \ldots, \omega^{p-1}\lambda$.

Exercise VIII.5.4 *Let $Z = \begin{pmatrix} R & A_1 \\ A_2 & -R \end{pmatrix}$, and suppose R commutes with A_1. Show that $\mathrm{tr}\, Z^k = 0$ if k is odd. Use this to show that λ is an eigenvalue of Z if and only if $-\lambda$ is.*

Exercise VIII.5.5 *Let ω be the primitive pth root of unity. If X, Y are two matrices such that $XY = \omega YX$, then $(X+Y)^p = X^p + Y^p$.*

Exercise VIII.5.6 *Let Z be a matrix of order $n = pr$ having a special form*

$$Z = \begin{pmatrix} R & A_1 & 0 & 0 & \cdots & 0 \\ 0 & \omega R & A_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & A_{p-1} \\ A_p & 0 & 0 & \cdots & \cdots & \omega^{p-1} R \end{pmatrix},$$

where R commutes with $A_1, A_2, \ldots, A_p$. Use the result of the preceding exercise to show that $\mathrm{tr}\; Z^k = 0$ if k is not an integral multiple of p. Use this

to show that if λ is an eigenvalue of Z, then so are $\omega\lambda, \omega^2\lambda, \cdots, \omega^{p-1}\lambda$. (This is true even when R commutes with $A_1, \ldots, A_{p-1}$.)

For brevity, an n-tuple will be called **p-Carrollian** if $n = pr$ and the elements of the tuple can be enumerated as

$$(\alpha_1, \ldots, \alpha_r, \omega\alpha_1, \ldots, \omega\alpha_r, \ldots, \omega^{p-1}\alpha_1, \ldots, \omega^{p-1}\alpha_r), \tag{VIII.32}$$

where ω is the primitive pth root of unity. We have seen above several examples of matrices whose eigenvalues are p-Carrollian.

Exercise VIII.5.7 *Let $s_k, 1 \leq k \leq n$ denote the elementary symmetric polynomials in n variables. If $(\alpha_1, \ldots, \alpha_n)$ is a Carrollian n-tuple written in the form (VIII.32), show that modulo a sign factor, we have*

$$s_k(\alpha_1, \ldots, \alpha_n) = \begin{cases} s_j(\alpha_1^p, \ldots, \alpha_r^p) & \text{if } k = jp \\ 0 & \text{if } k \neq jp. \end{cases}$$

Use this to show that if $\alpha_1, \ldots, \alpha_n$ are roots of the polynomial

$$f(z) = z^n + a_1 z^{n-1} + \cdots + a_n,$$

then $\alpha_1^p, \ldots, \alpha_r^p$ are roots of the polynomial

$$F(z) = z^r + a_p z^{r-1} + a_{2p} z^{r-2} + \cdots + a_{rp}.$$

Proposition VIII.5.8 *Let f, g be monic polynomials of degree n as in (VIII.10). Suppose $n = pr$ and the roots of f and g both are p-Carrollian. Let γ be as in (VIII.11). Then the roots of f and g can be labelled as $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ in such a way that*

$$\max_{1 \leq i \leq n} |\alpha_i^p - \beta_i^p| \leq 4 \left\{ \sum_{k=1}^{r} |a_{kp} - b_{kp}| \gamma^{p(r-k)} \right\}^{1/r}. \tag{VIII.33}$$

Proof. Use Theorem VIII.2.4 and Exercise VIII.5.7. ∎

Theorem VIII.5.9 *Let $n = pr$ and let A, B be two $n \times n$ matrices whose eigenvalues are p-Carrollian. Then*

$$d(\sigma(A^p), \sigma(B^p)) \leq 4\, c_{r,p}\, M^{p-1/r} \|A - B\|^{1/r}, \tag{VIII.34}$$

where $M = \max(\|A\|, \|B\|)$ and

$$c_{r,p} = \left\{ \sum_{k=1}^{r} kp \binom{rp}{kp} \right\}^{1/r}. \tag{VIII.35}$$

Proof. See Exercise VIII.2.6 and use the preceding proposition. ■

The two results above give bounds not on the distance between the roots themselves but on that between their pth powers. If all of them are outside a neighbourhood of zero, a bound on the distance between the roots can be obtained from this. This needs the following lemma.

Lemma VIII.5.10 *Let x, y be complex numbers such that $|x| \geq \rho$, $|y| \geq \rho$ and $|x^p - y^p| \leq C$. Then, for some k, $0 \leq k \leq p-1$,*

$$|x - \omega^k y| \leq \frac{C}{\rho^{p-1}}, \qquad \text{(VIII.36)}$$

where ω is the primitive pth root of unity.

Proof. Compare the coefficients of t in the identity

$$\prod_{k=0}^{p-1} [t - (x - \omega^k y)] = (-1)^p[(x-t)^p - y^p]$$

to see that

$$s_{p-1}(x - y,\ x - \omega y, \ldots, x - \omega^{p-1} y) = (-1)^{p-1} p x^{p-1}.$$

The right-hand side has modulus larger than $p\rho^{p-1}$ and the left-hand side is a sum of p terms. Hence, at least one of them should have modulus larger than ρ^{p-1}. So, there exists k, $0 \leq k \leq p-1$, such that

$$\prod_{\substack{j \neq k \\ j=0}}^{p-1} |x - \omega^j y| \geq \rho^{p-1}.$$

But $\prod_{j=0}^{p-1} |x - \omega^j y| = |x^p - y^p| \leq C$. This proves the lemma . ■

When $p = 2$, the inequality (VIII.36) can be strengthened. To see this note that

$$|x - y|^2 + |x + y|^2 = 2(|x|^2 + |y|^2) \geq 4\rho^2.$$

So, either $|x - y|$ or $|x + y|$ must be larger than $2^{1/2}\rho$. Consequently, one of them must be smaller than $C/2^{1/2}\rho$.

Thus if the eigenvalues of A and B are p-Carrollian and all have modulus larger than ρ, then $d(\sigma(A),\ \sigma(B))$ is bounded by C/ρ^{p-1}, where C is the quantity on the right-hand side of (VIII.34). When $p = 2$, this bound can be improved further to $C/\sqrt{2}\rho$. The major improvement over bounds obtained in Section 2 is that now the bounds involve $\|A - B\|^{1/r}$ instead of

$\|A-B\|^{1/n}$. For low values of p, the factors $c_{r,p}$ can be evaluated explicitly. For example, we have the combinatorial identity

$$\sum_{k=0}^{r} 2k\binom{n}{2k} = n2^{n-2} \quad \text{if } n = 2r.$$

VIII.6 Problems

Problem VIII.6.1. Let $f(z) = z^n + a_1 z^{n-1} + \cdots + a_n$ be a monic polynomial. Let $\mu_1, \dots, \mu_n$ be the numbers $|a_k|^{1/k}$, $1 \le k \le n$, rearranged in decreasing order. Show that all the roots of f are bounded (in absolute value) by $\mu_1 + \mu_2$. This is an improvement on the result of Lemma VIII.2.1.

Problem VIII.6.2. Fill in the details in the following alternate proof of Theorem VIII.3.1.

Let β be an eigenvalue of B but not of A. If $Bx = \beta x$, then

$$x = S(\beta I - D)^{-1} S^{-1}(B - A)x.$$

Hence,

$$\|x\| \le \text{cond}(S)\|B - A\| \; \|(\beta I - D)^{-1}\| \; \|x\|.$$

From this it follows that

$$\min_j |\beta - \alpha_j| \le \text{cond}(S)\|B - A\|.$$

Notice that this proof too relies only on those properties of $\|\cdot\|$ that are shared by many other norms (like the ones induced by the p-norms on $\mathbb{C}^n$). See the remark following Theorem VIII.3.1.

Problem VIII.6.3. Let B be any matrix with entries b_{ij}. The disks

$$D_i = \{z : |z - b_{ii}| \le \sum_{j \ne i} |b_{ij}|\}, \; 1 \le i \le n,$$

are called the **Gersgorin disks** of B. The **Gersgorin Disk Theorem** says that

$$\sigma(B) \subset \bigcup_{i=1}^{n} D_i,$$

and that any connected component of the set $\bigcup_i D_i$ contains as many eigenvalues of B as the number of disks that form this component.

The proof of this is outlined below.

Consider the vector norm $\|x\|_\infty = \max_{1\le i\le n} |x_i|$ on $\mathbb{C}^n$. The norm it induces on operators is

$$\|A\|_{\infty\to\infty} = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|.$$

Let D be the diagonal of B, and let $H = B - D$. Let β be an eigenvalue of B but not of D. Then

$$\beta I - B = \beta I - H - D = (\beta I - D)[I - (\beta I - D)^{-1}H].$$

Since $\beta I - B$ is not invertible, neither is the matrix in the square brackets. Hence,

$$1 \le \|(\beta I - D)^{-1}H\|_{\infty\to\infty}.$$

From this, the first part of the theorem follows. The second part follows from the continuity argument we have used often. Let $B(t) = D + tH$, $0 \le t \le 1$. Then $B(0) = D$, $B(1) = B$; the eigenvalues of $B(t)$ trace continuous curves that join the eigenvalues of D to those of B.

Note that the proof of the first part is very similar to that of Theorem VIII.3.3; in fact, it is a special case of the earlier one.

Problem VIII.6.4. Given any matrix A, we can find a unitary U such that

$$U^*AU = T = D + N,$$

where T is upper triangular, D is diagonal, and N is strictly upper triangular and, hence, nilpotent. Such a reduction is not unique. The **measure of nonnormality** of A is defined as

$$\Delta(A) = \inf \|N\|,$$

where the infimum is taken over all N that occur in the possible triangular forms of A given above.

Now let B be any other matrix, and let β be an eigenvalue of B but not of A. From (VIII.19) we have

$$\|(D - \beta I + N)^{-1}\|^{-1} \le \|A - B\|.$$

Show that

$$\begin{aligned}(D - \beta I + N)^{-1} &= [I + (D - \beta I)^{-1}N]^{-1}(D - \beta I)^{-1} \\ &= [I - (D - \beta I)^{-1}N + \{(D - \beta I)^{-1}N\}^2 \\ &\quad + \cdots + (-1)^{n-1}\{(D - \beta I)^{-1}N\}^{n-1}](D - \beta I)^{-1}.\end{aligned}$$

Let $\delta = \operatorname{dist}(\beta, \sigma(A))$. From this equation and the inequality before it conclude that

$$\|A - B\|^{-1} \le \frac{1}{\delta}\left\{1 + \frac{\Delta(A)}{\delta} + \left(\frac{\Delta(A)}{\delta}\right)^2 + \cdots + \left(\frac{\Delta(A)}{\delta}\right)^{n-1}\right\}.$$

Now show that

$$\frac{\left(\frac{\delta}{\Delta(A)}\right)^n}{1+\frac{\delta}{\Delta(A)}+\cdots+\left(\frac{\delta}{\Delta(A)}\right)^{n-1}} \le \frac{\|A-B\|}{\Delta(A)}.$$

This is **Henrici's Theorem.**

Let $f(t) = t^n/(1+t+\cdots+t^{n-1})$. Then $f(t)$ is close to t^n for small values of t, and to t for large values of t. Thus, when $\Delta(A)$ is close to 0, i.e., when A is close to being normal, the above bound leads to the asymptotic inequality

$$s(\sigma(B),\ \sigma(A)) \lesssim \|A-B\|.$$

In Theorem VI.3.3 we saw that if A is normal, then $s(\sigma(B),\ \sigma(A)) \le \|A-B\|$.

Problem VIII.6.5. Let ν be any norm on the space of matrices. The **ν-measure of nonnormality** of A is defined as

$$\Delta_\nu(A) = \inf\ \nu(N),$$

where N is as in Problem VIII.6.4. Suppose that the norm ν is such that $\|A\| \le \nu(A)$ for all A. Show that $\Delta(A)$ in Henrici's Theorem can be replaced by $\Delta_\nu(A)$.

Problem VIII.6.6. For the Hilbert-Schmidt norm $\|\cdot\|_2$, the measure of nonnormality satisfies the inequality

$$\Delta_2(A) \le \left(\frac{n^3-n}{12}\right)^{1/4} \|A^*A - AA^*\|_2^{1/2}$$

for every $n \times n$ matrix A. (The proof is a little intricate.)

Problem VIII.6.7. Let A have the Jordan canonical form $J = SAS^{-1}$. Let m be the size of the largest Jordan block in J. Let B be any other matrix. Show that for every eigenvalue β of B there is an eigenvalue α of A such that

$$\frac{|\beta-\alpha|^m}{(1+|\beta-\alpha|)^{m-1}} \le \|S(A-B)S^{-1}\|.$$

Problem VIII.6.8. Let A, B, Γ be as in Theorem VIII.3.4. Let $(A-B)x_j = \lambda_j x_j$, where the vectors x_j are orthonormal and the eigenvalues λ_j are indexed in such a way that $s_j := s_j(A-B) = |\lambda_j|$. Let y_j be the orthonormal vectors that satisfy the relations $(A-B)x_j = s_j y_j$. Note that $y_j = \pm x_j$. Note also that the difference of $A\Gamma - \Gamma B$ and $(A-B)\Gamma$

is skew-Hermitian. Use this to show that, for $1 \leq k \leq n$,

$$\begin{aligned} \text{Re} \sum_{j=1}^{k} \langle x_j, (A\Gamma - \Gamma B)y_j \rangle &= \text{Re} \sum_{j=1}^{k} \langle x_j, (A - B)\Gamma y_j \rangle \\ &= \sum_{j=1}^{k} s_j \langle y_j, \Gamma y_j \rangle \geq \gamma \sum_{j=1}^{k} s_j. \end{aligned}$$

Use this to give an alternate proof of Theorem VIII.3.4. (See Problem III.6.6.)

Problem VIII.6.9. Fill in the details in the following proof of Corollary VIII.3.7. Let $D = \Gamma - \gamma I$. Then

$$\begin{aligned} \|A\Gamma - \Gamma B\|_2^2 &= \|(AD - DB) + \gamma(A - B)\|_2^2 \\ &= \|AD - DB\|_2^2 + \gamma^2 \|A - B\|_2^2 \\ &\quad +2\gamma \text{ Re tr } (AD - DB)^*(A - B). \end{aligned}$$

So, it suffices to show that the last term is positive. This can be seen by writing

$$2 \text{ Re tr } (AD-DB)^*(A-B) = \text{tr } \{(AD-DB)^*(A-B)+(A-B)^*(AD-DB)\}$$

and then using cyclicity of the trace to reduce this to

$$\text{tr } D[(A - B)^*(A - B) + (A - B)(A - B)^*].$$

Problem VIII.6.10. (i) Let X be a contractive matrix; i.e., let $\|X\| \leq 1$. Show that there exist unitaries U and V such that $X = \frac{1}{2}(U+V)$. Use this to show that if D_1 and D_2 are real diagonal matrices, then

$$|||D_1X - XD_2||| \leq |||D_1^{\downarrow} - D_2^{\uparrow}|||$$

for every unitarily invariant norm. [(See (IV.62).]

(ii) Let $A = SD_1S^{-1}$, $B = TD_2T^{-1}$, where S and T are invertible matrices and D_1, D_2 are real diagonal matrices. Show that

$$|||A - B||| \leq \text{cond}(S)\text{cond}(T)|||\text{Eig}^{\downarrow}(A) - \text{Eig}^{\uparrow}(B)|||.$$

Problem VIII.6.11. Let A and B be any two diagonalisable matrices with eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_n$, respectively. Let $A = SD_1S^{-1}$, $B = TD_2T^{-1}$, where S and T are invertible matrices and D_1, D_2 are diagonal matrices. Show that

$$\|A - B\|_2 \leq \text{cond}(S)\text{cond}(T)\max_{\pi} \left(\sum_i |\lambda_i - \mu_{\pi(i)}|^2 \right)^{1/2},$$

where π varies over all permutations on n symbols. [See Theorem VI.4.1.]

Problem VIII.6.12. Let A be a Hermitian matrix with eigenvalues $\alpha_1, \ldots, \alpha_n$. Let B be any other matrix. For $1 \leq j \leq n$, let

$$D_j = \{z : |z - \alpha_j| \leq \|A - B\|, |\text{Im } z| \leq \|\text{Im}(A - B)\|\}.$$

The regions D_j are disks flattened on the top and bottom by horizontal lines. Show that the eigenvalues of B are contained in $\bigcup_j D_j$, and that each connected component of this set contains as many eigenvalues of A as of B.

Problem VIII.6.13. Let $\mathcal{R}$ be a real vector space whose elements are matrices with real eigenvalues. Show that the function $\sum_{j=1}^{k} \lambda_j^{\downarrow}(A)$ is a convex function of A on this space for $1 \leq k \leq n$. Show that the function $\sum_{j=1}^{k} \lambda_j^{\uparrow}(A)$ is concave on $\mathcal{R}$.

Problem VIII.6.14. If R_1 is invertible, then

$$\begin{pmatrix} R_1 & A_1 \\ A_2 & R_2 \end{pmatrix} = \begin{pmatrix} R_1 & 0 \\ A_2 & R_2 - A_2 R_1^{-1} A_1 \end{pmatrix} \begin{pmatrix} I & R_1^{-1} A_1 \\ 0 & I \end{pmatrix}.$$

Use this to show that if

$$Z = \begin{pmatrix} R & A_1 \\ A_2 & -R \end{pmatrix}$$

and R commutes with A_1, then Z and $-Z$ have the same eigenvalues. (Show that they have the same characteristic polynomials.) This gives another proof of the statement at the end of Exercise VIII.5.6, for $p = 2$. The same method works for $p > 2$. For instance, the case $p = 3$ is dealt with as follows. If R_1, R_2 are invertible, then

$$\begin{pmatrix} R_1 & A_1 & 0 \\ 0 & R_2 & A_2 \\ A_3 & 0 & R_3 \end{pmatrix}$$

$$= \begin{pmatrix} R_1 & 0 & 0 \\ 0 & R_2 & 0 \\ A_3 & -A_3 R_1^{-1} A_1 & R_3 + A_3 R_1^{-1} A_1 R_2^{-1} A_2 \end{pmatrix} \begin{pmatrix} I & R_1^{-1} A_1 & 0 \\ 0 & I & R_2^{-1} A_2 \\ 0 & 0 & I \end{pmatrix}.$$

Derive similar factorisations for $p > 3$, and use this to prove the statement at the end of Exercise VIII.5.6.

VIII.7 Notes and References

Many of the topics in this chapter have been presented earlier in R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman, 1987, and in G.W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, 1990. Some results that were proved after the publication of these books have, of course, been included here.

The first major results on perturbation of roots of polynomials were proved by A. Ostrowski, *Recherches sur la méthode de Gräffe et les zeros des polynômes et des series de Laurent*, Acta Math., 72 (1940), 99-257. See also Appendices A and B of his book *Solution of Equations and Systems of Equations*, Academic Press, 1960. Theorem VIII.2.2 is due to Ostrowski. Using this he proved an inequality weaker than (VIII.15); this had a factor $(2n-1)$ instead of 4. The argument used by him is the one followed in Exercise VIII.1.3.

Ostrowski was also the first to derive perturbation bounds for eigenvalues of arbitrary matrices in his paper *Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen*, Jber. Deut. Mat. - Verein, 60 (1957) 40-42. See also Appendix K of his book cited above. The inequality he proved involved the matrix norm $\|A\|_L = \frac{1}{n}\sum_{i,j}|a_{ij}|$, which is easy to compute but is not unitarily invariant. With this norm, his inequality is like the one in (VIII.4).

An inequality for $d(\sigma(A), \sigma(B))$ in terms of the unitarily invariant Hilbert-Schmidt norm was proved by R. Bhatia and K.K. Mukherjea, *On the rate of change of spectra of operators*, Linear Algebra Appl., 27 (1979) 147-157. They followed the approach in Exercise VIII.2.6 and, after a little tidying up, their result looks like (VIII.4) but with the larger norm $\|\cdot\|_2$ instead of $\|\cdot\|$. This approach was followed, to a greater success, in R. Bhatia and S. Friedland, *Variation of Grassmann powers and spectra*, Linear Algebra Appl., 40 (1981) 1-18. In this paper, the norm $\|\cdot\|$ was used and an inequality slightly weaker than (VIII.4) was proved.

An improvement of these inequalities in which $(2n-1)$ is replaced by n was made by L. Elsner, *On the variation of the spectra of matrices*, Linear Algebra Appl., 47 (1982) 127-138. The major insightful observation was that the Matching Theorem does not exploit the symmetry between the polynomials f and g, nor the matrices A and B, under consideration. Theorem VIII.1.1 is also due to L. Elsner, *An optimal bound for the spectral variation of two matrices*, Linear Algebra Appl., 71 (1985) 77-80.

The argument using Chebyshev polynomials, that we have employed in Sections VIII.1 and VIII.2, seems to have been first used by A. Schönhage, *Quasi-GCD computations*, J. Complexity, 1(1985) 118-137. (See Theorem 2.7 of this paper.) It was discovered independently by D. Phillips, *Improving spectral variation bounds with Chebyshev polynomials*, Linear Algebra Appl., 133 (1990) 165-173. Phillips proved a weaker inequality than (VIII.8)

with a factor 8 instead of 4.

This argument was somewhat simplified and used again by R. Bhatia, L. Elsner, and G. Krause, *Bounds for the variation of the roots of a polynomial and the eigenvalues of a matrix*, Linear Algebra Appl., 142 (1990) 195-209. Theorems VIII.1.5 and VIII.2.4 (and their proofs) have been taken from this paper. Using finer results from Chebyshev approximation, G. Krause has shown that the factor 4 occurring in these inequalities can be replaced by 3.08. See his paper *Bounds for the variation of matrix eigenvalues and polynomial roots*, Linear Algebra Appl., 208/209 (1994) 73-82. It was shown by Bhatia, Elsner, and Krause in the paper cited above that, in the inequality (VIII.15), the factor 4 cannot be replaced by anything smaller than 2.

Theorems VIII.3.1 and VIII.3.3 were proved in the very influential paper, F.L. Bauer and C.T. Fike, *Norms and exclusion theorems*, Numer. Math., 2 (1960) 137-141. See the discussion in Stewart and Sun, p. 177.

The basic idea behind results in Section VIII.3 from Theorem VIII.3.4 onwards is due to W. Kahan, *Inclusion theorems for clusters of eigenvalues of Hermitian matrices*, Technical Report, Computer Science Department, University of Toronto, 1967. Theorem VIII.3.4 for the special case of the operator norm is proved in this report. The inequality (VIII.23) is due to J.-G. Sun, *On the perturbation of the eigenvalues of a normal matrix*, Math. Numer. Sinica, 6(1984) 334-336. The ideas of Kahan's and Sun's proofs are outlined in Problems VIII.6.8 and VIII.6.9. Theorem VIII.3.4, in its generality, was proved in R. Bhatia, C. Davis, and F. Kittaneh, *Some inequalities for commutators and an application to spectral variation*, Aequationes Math., 41(1991) 70-78. The three corollaries were also proved there. These authors then used their commutator inequalities to derive weaker versions of Theorems VIII.3.9 and VIII.3.10; in all these, the square root in the inequalities (VIII.25) and (VIII.26) is missing. For the operator norm alone, the inequality (VIII.25) was proved by T.-X. Lu, *Perturbation bounds for eigenvalues of symmetrizable matrices*, Numerical Mathematics: a Journal of Chinese Universities, 16(1994) 177-185 (in Chinese). The inequalities (VIII.25)-(VIII.27) have been proved recently by R. Bhatia, F. Kittaneh and R.-C. Li, *Some inequalities for commutators and an application to spectral variation II*, Linear and Multilinear Algebra, to appear.

The inequality in Problem VIII.6.10 was proved in R. Bhatia, L. Elsner, and G. Krause, *Spectral variation bounds for diagonalisable matrices*, Preprint 94-098, SFB 343, University of Bielefeld. Example VIII.3.8 (and another example illustrating the same phenomenon for the trace norm) was constructed in this paper. The inequality in Problem VIII.6.11 was found by L. Elsner and S. Friedland, *Singular values, doubly stochastic matrices and applications*, Linear Algebra Appl., 220(1995) 161-169.

The results of Section VIII.4 were discovered by P. D. Lax, *Differential equations, difference equations and matrix theory*, Comm. Pure Appl. Math., 11(1958) 175-194. Lax was motivated by the theory of linear partial

differential equations of hyperbolic type, and his proofs used techniques from this theory. The paper of Lax was followed by one by H.F. Weinberger, *Remarks on the preceding paper of Lax*, Comm. Pure Appl. Math., 11 (1958) 195-196. He gave simple matrix theoretic proofs of these theorems, which we have reproduced here. L. Garding later pointed out that these are special cases of his results for hyperbolic polynomials that appeared in his papers *Linear hyperbolic partial differential equations with constant coefficients*, Acta Math., 84(1951) 1-62, and *An inequality for hyperbolic polynomials*, J. Math. Mech., 8(1959) 957-966. A characterisation of the kind of spaces $\mathcal{R}$ discussed in Section VIII.4 was given by H. Wielandt, *Lineare Scharen von Matrizen mit reellen Eigenwerten*, Math. Z., 53(1950) 219-225.

It was observed by R. Bhatia, *On the rate of change of spectra of operators II*, Linear Algebra Appl., 36 (1981) 25-32, that better perturbation bounds can be obtained for matrices whose eigenvalues occur in pairs $\pm\lambda$. This was carried further in the paper *Symmetries and variation of spectra*, Canadian J. Math., 44 (1992) 1155-1166, by R. Bhatia and L. Elsner, who considered matrices whose eigenvalues are p-Carrollian. See also the paper by R. Bhatia and L. Elsner, *The q-binomial theorem and spectral symmetry*, Indag. Math., N.S., 4(1993) 11-16. The material in Section VIII.5 is taken from these three papers.

The bound in Problem VIII.6.1 is due to Lagrange. There are several interesting and useful bounds known for the roots of a polynomial. Since the roots of a polynomial are the eigenvalues of its companion matrix, some of these bounds can be proved by using bounds for eigenvalues. An interesting discussion may be found in Horn and Johnson, *Matrix Analysis*, pages 316-319.

The Gersgorin Disk Theorem was proved in S.A. Gersgorin, *Über die Abrenzung der Eigenwerte einer Matrix*, Izv. Akad. Nauk SSSR, Ser. Fiz. - Mat., 6(1931) 749-754. A matrix is called **diagonally dominant** if $|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad 1 \leq i \leq n$. Every diagonally dominant matrix is nonsingular. Gersgorin's Theorem is a corollary. This theorem is applied to the study of several perturbation problems in J.H. Wilkinson, *The Algebraic Eigenvalue Problem.* A comprehensive discussion is also given in Horn and Johnson, *Matrix Analysis.*

The results of Problems VIII.6.4, VIII.6.5, and VIII.6.6 are due to P. Henrici, *Bounds for iterates, inverses, spectral variation and fields of values of nonnormal matrices*, Numer. Math., 4 (1962) 24-39. Several other very interesting results that involve the measure of nonnormality are proved in this paper. For example, we know that the numerical range $W(A)$ of a matrix A contains the convex hull $H(A)$ of the eigenvalues of A, and that the two sets are equal if A is normal. Henrici gives a bound for the distance between the boundaries of $H(A)$ and $W(A)$ in terms of the measure of nonnormality of A.

There are several different ways to measure the nonnormality of a matrix. Problem VIII.6.6 relates two such measures by an inequality. The relations between several different measures of nonnormality are discussed in L. Elsner and M.H.C. Paardekooper, *On measures of nonnormality of matrices*, Linear Algebra Appl., 92(1987) 107-124.

Is a *nearly normal* matrix near to an (exactly) normal matrix? More precisely, for every $\varepsilon > 0$, does there exist a $\delta > 0$ such that if $\|A^*A - AA^*\| < \delta$ then there exists a normal B such that $\|A - B\| < \varepsilon$? The existence of such a δ for each fixed dimension n was shown by C. Pearcy and A. Shields, *Almost commuting matrices*, J. Funct. Anal., 33(1979) 332-338. The problem of finding a δ depending only on ε but not on the dimension n is linked to several important questions in the theory of operator algebras. This has been shown to have an affirmative solution in a recent paper: H. Lin, *Almost commuting selfadjoint matrices and applications*, preprint, 1995. No explicit formula for δ is given in this paper. In an infinite-dimensional Hilbert space, the answer to this question is in the negative because of index obstructions.

The inequality in Problem VIII.6.7 was proved in W. Kahan, B.N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal. 19(1982) 470-484.

The inequality in Problem VIII.6.12 was proved by W. Kahan, *Spectra of nearly Hermitian matrices*, Proc. Amer. Math. Soc., 48(1975) 11-17.

For 2×2 block-matrices, the idea of the argument in Problem VIII.6.14 is due to M.D. Choi, *Almost commuting matrices need not be nearly commuting*, Proc. Amer. Math. Soc. 102(1988) 529-533. This was extended to higher order block-matrices by R. Bhatia and L. Elsner, *Symmetries and variation of spectra*, cited above.

IX

A Selection of Matrix Inequalities

In this chapter we will prove several inequalities for matrices. From the vast collection of such inequalities, we have selected a few that are simple and widely useful. Though they are of different kinds, their proofs have common ingredients already familiar to us from earlier chapters.

IX.1 Some Basic Lemmas

If A and B are any two matrices, then AB and BA have the same eigenvalues. (See Exercise I.3.7.) Hence, if $f(A)$ is any function on the space of matrices that depends only on the eigenvalues of A, then $f(AB) = f(BA)$. Examples of such functions are the spectral radius, the trace, and the determinant. If A is normal, then the spectral radius $\operatorname{spr}(A)$ is equal to $\|A\|$. Using this, we can prove the following two useful propositions.

Proposition IX.1.1 *Let A, B be any two matrices such that the product AB is normal. Then, for every unitarily invariant norm, we have*

$$|||AB||| \leq |||BA|||. \tag{IX.1}$$

Proof. For the operator norm this is an easy consequence of the two facts mentioned above; we have

$$\|AB\| = \operatorname{spr}(AB) = \operatorname{spr}(BA) \leq \|BA\|.$$

The general case needs more argument. Since AB is normal, $s_j(AB) = |\lambda_j(AB)|$, where $|\lambda_1(AB)| \geq \cdots \geq |\lambda_n(AB)|$ are the eigenvalues of AB

arranged in decreasing order of magnitude. But $|\lambda_j(AB)| = |\lambda_j(BA)|$. By Weyl's Majorant Theorem (Theorem II.3.6), the vector $|\lambda(BA)|$ is weakly majorised by the vector $s(BA)$. Hence we have the weak majorisation $s(AB) \prec_w s(BA)$. From this the inequality (IX.1) follows. ∎

Proposition IX.1.2 *Let A, B be any two matrices such that the product AB is Hermitian. Then, for every unitarily invariant norm, we have*

$$|||AB||| \leq |||\mathrm{Re}(BA)|||. \tag{IX.2}$$

Proof. The eigenvalues of BA, being the same as the eigenvalues of the Hermitian matrix AB, are all real. So, by Proposition III.5.3, we have the majorisation $\lambda(BA) \prec \lambda(\mathrm{Re}\ BA)$. From this we have the weak majorisation $|\lambda(BA)| \prec_w |\lambda(\mathrm{Re}\ BA)|$. (See Examples II.3.5.) The rest of the argument is the same as in the proof of the preceding proposition. ∎

Some of the inequalities proved in this chapter involve the matrix exponential. An extremely useful device in proving such results is the following theorem.

Theorem IX.1.3 *(The Lie Product Formula) For any two matrices A, B,*

$$\lim_{m\to\infty} \left(\exp \frac{A}{m} \exp \frac{B}{m}\right)^m = \exp(A+B). \tag{IX.3}$$

Proof. For any two matrices X, Y, and for $m = 1, 2, \ldots$, we have

$$X^m - Y^m = \sum_{j=0}^{m-1} X^{m-1-j}(X-Y)Y^j.$$

Using this we obtain

$$\|X^m - Y^m\| \leq m\, M^{m-1}\|X - Y\|, \tag{IX.4}$$

where $M = \max(\|X\|,\ \|Y\|)$.

Now let $X_m = \exp(\frac{A+B}{m})$, $Y_m = \exp\ \frac{A}{m}\ \exp\ \frac{B}{m}$, $m = 1, 2, \ldots$. Then $\|X_m\|$ and $\|Y_m\|$ both are bounded above by $\exp\left(\frac{\|A\|+\|B\|}{m}\right)$. From the power series expansion for the exponential function, we see that

$$\begin{aligned} X_m - Y_m &= 1 + \frac{A+B}{m} + \frac{1}{2}\left(\frac{A+B}{m}\right)^2 + \cdots \\ &\quad - \left\{\left[1 + \frac{A}{m} + \frac{1}{2}\left(\frac{A}{m}\right)^2 + \cdots\right]\left[1 + \frac{B}{m} + \frac{1}{2}\left(\frac{B}{m}\right)^2 + \cdots\right]\right\} \\ &= O\left(\frac{1}{m^2}\right) \quad \text{for large } m. \end{aligned}$$

Hence, using the inequality (IX.4), we see that

$$\|X_m^m - Y_m^m\| \leq m \exp(\|A\| + \|B\|) O\left(\frac{1}{m^2}\right).$$

This goes to zero as $m \to \infty$. But $X_m^m = \exp(A+B)$ for all m. Hence, $\lim_{m\to\infty} Y_m^m = \exp(A+B)$. This proves the theorem. ■

The reader should compare the inequality (IX.4) with the inequality in Problem I.6.11.

Exercise IX.1.4 *Show that for any two matrices A, B*

$$\lim_{m\to\infty}\left(\exp\frac{B}{2m}\ \exp\frac{A}{m}\ \exp\frac{B}{2m}\right)^m = \exp(A+B).$$

Exercise IX.1.5 *Show that for any two matrices A, B*

$$\lim_{t\to 0}\left(\exp\frac{tB}{2}\ \exp tA\ \exp\frac{tB}{2}\right)^{1/t} = \exp(A+B).$$

IX.2 Products of Positive Matrices

In this section we prove some inequalities for the norm, the spectral radius, and the eigenvalues of the product of two positive matrices.

Theorem IX.2.1 *Let A, B be positive matrices. Then*

$$\|A^s B^s\| \leq \|AB\|^s, \quad \text{for} \quad 0 \leq s \leq 1. \tag{IX.5}$$

Proof. Let

$$D = \{s : 0 \leq s \leq 1,\ \|A^s B^s\| \leq \|AB\|^s\}.$$

Then D is a closed subset of $[0,1]$ and contains the points 0 and 1. So, to prove the theorem, it suffices to prove that if s and t are in D then so is $\frac{s+t}{2}$. We have

$$\begin{aligned} \|A^{\frac{s+t}{2}} B^{\frac{s+t}{2}}\|^2 &= \|B^{\frac{s+t}{2}} A^{s+t} B^{\frac{s+t}{2}}\| = \operatorname{spr}(B^{\frac{s+t}{2}} A^{s+t} B^{\frac{s+t}{2}}) \\ &= \operatorname{spr}(B^s A^{s+t} B^t) \leq \|B^s A^{s+t} B^t\| \\ &\leq \|B^s A^s\|\ \|A^t B^t\| = \|A^s B^s\|\ \|A^t B^t\|. \end{aligned}$$

At the first step we used the relation $\|T\|^2 = \|T^*T\|$, and at the last step the relation $\|T^*\| = \|T\|$ for all T. If s, t are in D, this shows that

$$\|A^{\frac{s+t}{2}} B^{\frac{s+t}{2}}\| \leq \|AB\|^{(s+t)/2},$$

and this proves the theorem. ∎

An equivalent formulation of the above theorem is given below, with another proof that is illuminating.

Theorem IX.2.2 *If A, B are positive matrices with $\|AB\| \leq 1$, then $\|A^s B^s\| \leq 1$ for $0 \leq s \leq 1$.*

Proof. We can assume that $A > 0$. The general case follows from this by a continuity argument. We then have the chain of implications

$$\begin{aligned} \|AB\| \leq 1 \quad &\Rightarrow \quad \|AB^2A\| \leq 1 \Rightarrow AB^2A \leq I \\ &\Rightarrow \quad B^2 \leq A^{-2} \text{ (by Lemma V.1.5)} \\ &\Rightarrow \quad B^{2s} \leq A^{-2s} \text{ (by Theorem V.1.9)} \\ &\Rightarrow \quad A^s B^{2s} A^s \leq I \text{ (by Lemma V.1.5)} \\ &\Rightarrow \quad \|A^s B^{2s} A^s\| \leq 1 \Rightarrow \|A^s B^s\| \leq 1. \end{aligned}$$

∎

Another equivalent formulation is the following theorem.

Theorem IX.2.3 *Let A, B be positive matrices. Then*

$$\|AB\|^t \leq \|A^t B^t\|, \quad \textit{for} \quad t \geq 1. \tag{IX.6}$$

Proof. From Theorem IX.2.1, we have $\|A^{1/t}\, B^{1/t}\| \leq \|AB\|^{1/t}$ for $t \geq 1$. Replace A, B by $A^t,\ B^t$, respectively. ∎

Exercise IX.2.4 *Let A, B be positive matrices. Then*

(i) $\|A^{1/t}\, B^{1/t}\|^t$ is a monotonically decreasing function of t on $(0, \infty)$.

(ii) $\|A^t B^t\|^{1/t}$ is a monotonically increasing function of t on $(0, \infty)$.

In Section 5 we will see that the inequalities (IX.5) and (IX.6) are, in fact, valid for all unitarily invariant norms.

Results akin to the ones above can be proved for the spectral radius in place of the norm. This is done below.

If A and B are positive, the eigenvalues of AB are positive. (They are the same as the eigenvalues of the positive matrix $A^{1/2}BA^{1/2}$.) If T is any matrix with positive eigenvalues, we will enumerate its eigenvalues as $\lambda_1(T) \geq \lambda_2(T) \geq \cdots \geq \lambda_n(T) \geq 0$. Thus $\lambda_1(T)$ is equal to the spectral radius spr (T).

Theorem IX.2.5 *If A, B are positive matrices with $\lambda_1(AB) \leq 1$, then $\lambda_1(A^s B^s) \leq 1$ for $0 \leq s \leq 1$.*

Proof. As in the proof of Theorem IX.2.2, we can assume that $A > 0$. We then have the chain of implications

$$\begin{aligned}\lambda_1(AB) \le 1 \;&\Rightarrow\; \lambda_1(A^{1/2}BA^{1/2}) \le 1 \Rightarrow A^{1/2}BA^{1/2} \le I\\ &\Rightarrow\; B \le A^{-1} \Rightarrow B^s \le A^{-s} \Rightarrow A^{s/2}B^sA^{s/2} \le I\\ &\Rightarrow\; \lambda_1(A^{s/2}B^sA^{s/2}) \le 1 \Rightarrow \lambda_1(A^sB^s) \le 1.\end{aligned}$$

This proves the theorem. ∎

It should be noted that all implications in this proof and that of Theorem IX.2.2 are reversible with one exception: if $A \ge B \ge 0$, then $A^s \ge B^s$ for $0 \le s \le 1$, but the converse is not true.

Theorem IX.2.6 *Let A, B be positive matrices. Then*

$$\lambda_1(A^sB^s) \le \lambda_1^s(AB), \quad \textit{for} \quad 0 \le s \le 1. \tag{IX.7}$$

Proof. Let $\lambda_1(AB) = \alpha^2$. If $\alpha \ne 0$, we have $\lambda_1(\frac{A}{\alpha}\,\frac{B}{\alpha}) = 1$. So, by Theorem IX.2.5, $\lambda_1(A^sB^s) \le \alpha^{2s} = \lambda_1^s(AB)$.

If $\alpha = 0$, we have $\lambda_1(A^{1/2}BA^{1/2}) = 0$, and hence $A^{1/2}BA^{1/2} = 0$. From this it follows that the range of A is contained in the kernel of B. But then $A^{s/2}B^sA^{s/2} = 0$, and hence, $\lambda_1(A^sB^s) = 0$. ∎

Exercise IX.2.7 *Let A, B be positive matrices. Show that*

$$\lambda_1^t(AB) \le \lambda_1(A^tB^t), \quad \textit{for} \quad t \ge 1. \tag{IX.8}$$

Exercise IX.2.8 *Let A, B be positive matrices. Show that*

(i) $[\lambda_1(A^{1/t}B^{1/t})]^t$ *is a monotonically decreasing function of t on* $(0, \infty)$.

(ii) $[\lambda_1(A^tB^t)]^{1/t}$ *is a monotonically increasing function of t on* $(0, \infty)$.

Using familiar arguments involving antisymmetric tensor products, we can now obtain stronger results.

Theorem IX.2.9 *Let A, B be positive matrices. Then, for $0 < t \le u < \infty$, we have the weak majorisation*

$$\lambda^{1/t}(A^tB^t) \prec_w \lambda^{1/u}(A^uB^u). \tag{IX.9}$$

Proof. For $k = 1, 2, \ldots, n$, consider the operators $\wedge^k A$ and $\wedge^k B$. The result of Exercise IX.2.8(ii) applied to these operators in place of A, B yields the inequalities

$$\prod_{j=1}^{k} \lambda_j^{1/t}(A^tB^t) \le \prod_{j=1}^{k} \lambda_j^{1/u}(A^uB^u) \tag{IX.10}$$

for $k = 1, 2, \ldots, n$. The assertion of II.3.5(vii) now leads to the majorisation (IX.9). ∎

Theorem IX.2.10 *Let A, B be positive matrices. Then for every unitarily invariant norm we have*

$$|||B^tA^tB^t||| \le |||(BAB)^t|||, \; for\, 0 \le t \le 1, \quad \text{(IX.11)}$$
$$|||(BAB)^t||| \le |||B^tA^tB^t|||, \; for\, t \ge 1. \quad \text{(IX.12)}$$

Proof. We have

$$\|B^tA^tB^t\| = \|(A^{t/2}B^t)^*(A^{t/2}B^t)\| = \|A^{t/2}B^t\|^2 \le \|A^{1/2}B\|^{2t},$$

for $0 \le t \le 1$, by Theorem IX.2.1. So

$$\|B^tA^tB^t\| \le \|BAB\|^t, \text{ for } 0 \le t \le 1.$$

This is the same as saying that

$$s_1(B^tA^tB^t) \le s_1^t(BAB).$$

Replacing A and B by their antisymmetric tensor powers, we obtain, for $1 \le k \le n$,

$$\prod_{j=1}^k s_j(B^tA^tB^t) \le \prod_{j=1}^k s_j^t(BAB).$$

By the argument used in the preceding theorem, this gives the majorisation

$$s(B^tA^tB^t) \prec_w s([BAB]^t),$$

which gives the inequality (IX.11).

The inequality (IX.12) is proved in exactly the same way. ■

Exercise IX.2.11 *Derive (as a special case of the above theorem) the following inequality of Araki-Lieb-Thirring. Let A, B be positive matrices, and let s, t be positive real numbers with $t \ge 1$. Then*

$$\operatorname{tr}[(B^{1/2}AB^{1/2})^{st}] \le \operatorname{tr}[(B^{t/2}A^tB^{t/2})^s]. \quad \text{(IX.13)}$$

IX.3 Inequalities for the Exponential Function

For every complex number z, we have $|e^z| = |e^{\operatorname{Re} z}|$. Our first theorem is a matrix version of this.

Theorem IX.3.1 *Let A be any matrix. Then*

$$|||e^A||| \le |||e^{\operatorname{Re} A}||| \quad \text{(IX.14)}$$

for every unitarily invariant norm.

Proof. For each positive integer m, we have $\|A^m\| \le \|A\|^m$. This is the same as saying that $s_1^2(A^m) \le s_1^{2m}(A)$ or $s_1(A^{*m}A^m) \le s_1^m(A^*A)$. Replacing A by $\wedge^k A$, we obtain for $1 \le k \le n$

$$\prod_{j=1}^{k} s_j(A^{*m}A^m) \le \prod_{j=1}^{k} s_j([A^*A]^m).$$

Now, if we replace A by $e^{A/m}$, we obtain

$$\prod_{j=1}^{k} s_j(e^{A^*}e^{A}) \le \prod_{j=1}^{k} s_j([e^{A^*/m}e^{A/m}]^m).$$

Letting $m \to \infty$, and using the Lie Product Formula, we obtain

$$\prod_{j=1}^{k} s_j(e^{A^*}e^{A}) \le \prod_{j=1}^{k} s_j(e^{A^*+A}).$$

Taking square roots, we get

$$\prod_{j=1}^{k} s_j(e^{A}) \le \prod_{j=1}^{k} s_j(e^{\mathrm{Re}\,A}).$$

This gives the majorisation

$$s(e^A) \prec_w s(e^{\mathrm{Re}\,A})$$

(see II.3.5(vii)), and hence the inequality (IX.14). ■

It is easy to construct an example of a 2×2 matrix A, for which $|||e^A|||$ and $|||e^{\mathrm{Re}\,A}|||$ are not equal.

Our next theorem is valid for a large class of functions. It will be convenient to give this class a name.

Definition IX.3.2 *A continuous complex-valued function f on the space of matrices will be said to belong to the class $\mathcal{T}$ if it satisfies the following two properties:*

(i) $f(XY) = f(YX)$ for all X, Y.

(ii) $|f(X^{2m})| \le f([XX^*]^m)$ for all X, and for $m = 1, 2, \ldots$.

Exercise IX.3.3 *(i) The functions trace and determinant are in $\mathcal{T}$.*

(ii) For every k, $1 \le k \le n$, the function $\varphi_k(X) = \mathrm{tr} \wedge^k X$ is in $\mathcal{T}$. (These are the coefficients in the characteristic polynomial of X.)

(iii) Let $\lambda_j(X)$ denote the eigenvalues of X arranged so that $|\lambda_1(X)| \geq |\lambda_2(X)| \geq \cdots \geq |\lambda_n(X)|$. Then, for $1 \leq k \leq n$, the function $f_k(X) = \prod_{j=1}^{k} \lambda_j(X)$ is in $\mathcal{T}$, and so is the function $|f_k(X)|$. [Hint: Use Theorem II.3.6.]

(iv) For $1 \leq k \leq n$, the function $g_k(X) = \sum_{j=1}^{k} |\lambda_j(X)|$ is in $\mathcal{T}$.

(v) Every symmetric gauge function of the numbers $|\lambda_1(X)|, \ldots, |\lambda_n(X)|$ is in $\mathcal{T}$.

Exercise IX.3.4 *(i) If f is any complex valued function on the space of matrices that satisfies the condition (ii) in Definition IX.3.2, then $f(A) \geq 0$ if $A \geq 0$. In particular, $f(e^A) \geq 0$ for every Hermitian matrix A.*

(ii) If f satisfies both conditions (i) and (ii) in Definition IX.3.2, then $f(AB) \geq 0$ if A and B are both positive. In particular, $f(e^A e^B) \geq 0$ if A and B are Hermitian.

The principal result about the class $\mathcal{T}$ is the following.

Theorem IX.3.5 *Let f be a function in the class $\mathcal{T}$. Then for all matrices A, B, we have*

$$|f(e^{A+B})| \leq f(e^{\text{Re } A} e^{\text{Re } B}). \tag{IX.15}$$

Proof. For each positive integer m, we have for all X, Y

$$\begin{aligned} |f([XY]^{2^m})| &\leq f([(XY)(XY)^*]^{2^{m-1}}) \\ &= f([XYY^*X^*]^{2^{m-1}}) \\ &= f([X^*XYY^*]^{2^{m-1}}). \end{aligned}$$

Here, the inequality at the first step is a consequence of the property (ii), and the equality at the last step is a consequence of the property (i) of functions in $\mathcal{T}$. Repeat this argument to obtain

$$\begin{aligned} |f([XY]^{2^m})| &\leq f([(X^*X)^2(YY^*)^2]^{2^{m-2}}) \\ &\leq f([X^*X]^{2^{m-1}}[YY^*]^{2^{m-1}}). \end{aligned}$$

Now let A, B be any two matrices. Put $X = e^{A/2^m}$ and $Y = e^{B/2^m}$ in the above inequality to obtain

$$|f([e^{A/2^m} e^{B/2^m}]^{2^m})| \leq f([e^{A^*/2^m} e^{A/2^m}]^{2^{m-1}} [e^{B/2^m} e^{B^*/2^m}]^{2^{m-1}}).$$

Now let $m \to \infty$. Then, by the continuity of f and the Lie Product Formula, we can conclude from the above inequality that

$$|f(e^{A+B})| \leq f(e^{(A+A^*)/2} e^{(B+B^*)/2}) = f(e^{\text{Re } A} e^{\text{Re } B}).$$

■

Corollary IX.3.6 *Let f be a function in the class $\mathcal{T}$. Then*

$$|f(e^A)| \leq f(e^{\operatorname{Re} A}), \quad \textit{for all } A, \tag{IX.16}$$

and

$$0 \leq f(e^{A+B}) \leq f(e^A e^B), \quad \textit{for Hermitian } A, B. \tag{IX.17}$$

Particularly noteworthy is the following special consequence.

Theorem IX.3.7 *Let A, B be any two Hermitian matrices. Then*

$$|||e^{A+B}||| \leq |||e^A e^B||| \tag{IX.18}$$

for every unitarily invariant norm.

Proof. Use (IX.17) for the special functions in Exercise IX.3.3(iv). This gives the majorisation

$$\lambda(e^{A+B}) \prec_w \lambda(e^A e^B).$$

But $\lambda(e^{A+B}) = s(e^{A+B})$ and $\lambda(e^A e^B) \prec_w s(e^A e^B)$. Hence

$$s(e^{A+B}) \prec_w s(e^A e^B).$$

This proves the theorem. ■

Choosing $f(X) = \operatorname{tr} X$, we get from (IX.17) the famous **Golden-Thompson inequality**: for Hermitian A, B we have

$$\operatorname{tr}(e^{A+B}) \leq \operatorname{tr}(e^A e^B). \tag{IX.19}$$

Exercise IX.3.8 *Let A, B be Hermitian matrices. Show that for every unitarily invariant norm*

$$\left|\left|\left| \exp \frac{tB}{2} \exp tA \exp \frac{tB}{2} \right|\right|\right|^{1/t}$$

decreases to $|||\exp(A+B)|||$ as $t \downarrow 0$. As a special consequence of this we have a stronger version of the Golden-Thompson inequality:

$$\operatorname{tr} \exp(A+B) \leq \operatorname{tr}\left(\exp \frac{tB}{2} \exp tA \exp \frac{tB}{2}\right) \quad \textit{for all} \quad t > 0.$$

[Use Theorem IX.2.10 and Exercise IX.1.5.]

IX.4 Arithmetic-Geometric Mean Inequalities

The classical arithmetic-geometric mean inequality for numbers says that $\sqrt{ab} \leq \frac{1}{2}(a+b)$ for all positive numbers a, b. From this we see that for complex numbers a, b we have $|\bar{a}b| \leq \frac{1}{2}(|a|^2 + |b|^2)$. In this section we obtain some matrix versions of this inequality. Several corollaries of these inequalities are derived in this and later sections.

Lemma IX.4.1 *Let Y_1, Y_2 be any two positive matrices, and let $Y = Y_1 - Y_2$. Let $Y = Y^+ - Y^-$ be the Jordan decomposition of the Hermitian matrix Y. Then, for $j = 1, 2, \ldots, n$,*

$$\lambda_j(Y^+) \leq \lambda_j(Y_1), \quad \lambda_j(Y^-) \leq \lambda_j(Y_2).$$

(See Section IV.3 for the definition of the Jordan decomposition.)

Proof. Suppose $\lambda_j(Y)$ is nonnegative for $j = 1, \ldots, p$ and negative for $j = p+1, \ldots, n$. Then $\lambda_j(Y^+)$ is equal to $\lambda_j(Y)$ if $j = 1, \ldots, p$, and is zero for $j = p+1, \ldots, n$.

Since $Y_1 = Y + Y_2 \geq Y$, we have $\lambda_j(Y_1) \geq \lambda_j(Y)$ for all j, by Weyl's Monotonicity Principle. Hence, $\lambda_j(Y_1) \geq \lambda_j(Y^+)$ for all j.

Since $Y_2 = Y_1 - Y \geq -Y$, we have $\lambda_j(Y_2) \geq \lambda_j(-Y)$ for all j. But $\lambda_j(-Y) = \lambda_j(Y^-)$ for $j = 1, \ldots, n-p$ and $\lambda_j(-Y) = 0$ for $j > n-p$. Hence, $\lambda_j(Y_2) \geq \lambda_j(Y^-)$ for all j. ■

Theorem IX.4.2 *Let A, B be any two matrices. Then*

$$s_j(A^*B) \leq \frac{1}{2} s_j(AA^* + BB^*) \tag{IX.20}$$

for $1 \leq j \leq n$.

Proof. Let X be the $2n \times 2n$ matrix $X = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}$. Then

$$XX^* = \begin{pmatrix} AA^* + BB^* & 0 \\ 0 & 0 \end{pmatrix}, \quad X^*X = \begin{pmatrix} A^*A & A^*B \\ B^*A & B^*B \end{pmatrix}.$$

The off-diagonal part of X^*X can be written as

$$Y = \begin{pmatrix} 0 & A^*B \\ B^*A & 0 \end{pmatrix} = \frac{1}{2}\{X^*X - U(X^*X)U^*\},$$

where U is the unitary matrix $\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$. Note that both of the matrices in the braces above are positive. Hence, by the preceding lemma,

$$\lambda_j(Y^+) \leq \frac{1}{2}\lambda_j(X^*X), \quad \lambda_j(Y^-) \leq \frac{1}{2}\lambda_j(X^*X).$$

But X^*X and XX^* have the same eigenvalues. Hence, both $\lambda_j(Y^+)$ and $\lambda_j(Y^-)$ are bounded above by $\frac{1}{2}\lambda_j(AA^*+BB^*)$. Now note that, by Exercise II.1.15, the eigenvalues of Y are the singular values of A^*B together with their negatives. Hence we have

$$s_j(A^*B) \leq \frac{1}{2}s_j(AA^* + BB^*).$$

■

Corollary IX.4.3 *Let A, B be any two matrices. Then there exists a unitary matrix U such that*

$$|A^*B| \leq \frac{1}{2}U(AA^* + BB^*)U^*. \tag{IX.21}$$

Corollary IX.4.4 *Let A, B be any two matrices. Then*

$$|||A^*B||| \leq \frac{1}{2}|||AA^* + BB^*||| \tag{IX.22}$$

for every unitarily invariant norm.

The particular position of the stars in (IX.20), (IX.21), and (IX.22) is not an accident. If we have

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix},$$

then $s_1(AB) = \sqrt{2}$, but $\frac{1}{2}s_1(AA^* + BB^*) = 1$.

The presence of the unitary U in (IX.21) is also essential: it cannot be replaced by the identity matrix even when A, B are Hermitian. This is illustrated by the example

$$A = B = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

A considerable strengthening of the inequality (IX.22) is given in the theorem below.

Theorem IX.4.5 *For any three matrices A, B, X, we have*

$$|||A^*XB||| \leq \frac{1}{2}|||AA^*X + XBB^*||| \tag{IX.23}$$

for every unitarily invariant norm.

Proof. First consider the special case when A, B, X are Hermitian and $A = B$. Then AXA is also Hermitian. So, by Proposition IX.1.2,

$$|||AXA||| \leq |||\mathrm{Re}(XA^2)||| = \frac{1}{2}|||A^2X + XA^2|||, \tag{IX.24}$$

which is just the desired inequality in this special case.

Next consider the more general situation, when A and B are Hermitian and X is any matrix. Let

$$T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix}.$$

Then, by the special case considered above,

$$|||TYT||| \leq \frac{1}{2} |||T^2Y + YT^2|||. \tag{IX.25}$$

Multiplying out the block-matrices, one sees that

$$\begin{aligned} TYT &= \begin{pmatrix} 0 & AXB \\ BX^*A & 0 \end{pmatrix}, \\ T^2Y + YT^2 &= \begin{pmatrix} 0 & A^2X + XB^2 \\ B^2X^* + X^*A^2 & 0 \end{pmatrix}. \end{aligned}$$

Hence, we obtain from (IX.25) the inequality

$$|||AXB||| \leq \frac{1}{2} |||A^2X + XB^2|||. \tag{IX.26}$$

Finally, let A, B, X be any matrices. Let $A = A_1U, \;\; B = B_1V$ be polar decompositions of A and B. Then

$$AA^*X + XBB^* = A_1^2X + XB_1^2,$$

while

$$|||A^*XB||| = |||UA_1XB_1V||| = |||A_1XB_1|||.$$

So, the theorem follows from the inequality (IX.26). ∎

Exercise IX.4.6 *Another proof of the theorem can be obtained as follows. First prove the inequality (IX.24) for Hermitian A and X. Then, for arbitrary A, B and X, let T and Y be the matrices*

$$T = \begin{pmatrix} 0 & 0 & A & 0 \\ 0 & 0 & 0 & B \\ A^* & 0 & 0 & 0 \\ 0 & B^* & 0 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & X & 0 & 0 \\ X^* & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and apply the special case to them.

Exercise IX.4.7 *Construct an example to show that the inequality (IX.20) cannot be strengthened in the way that (IX.23) strengthens (IX.22).*

When A, B are both positive, we can prove a result stronger than (IX.26). This is the next theorem.

Theorem IX.4.8 *Let A, B be positive matrices and let X be any matrix. Then, for each unitarily invariant norm, the function*

$$f(t) = |||A^{1+t}XB^{1-t} + A^{1-t}XB^{1+t}||| \tag{IX.27}$$

is convex on the interval $[-1, 1]$ and attains its minimum at $t = 0$.

Proof. Without loss of generality, we may assume that $A > 0$ and $B > 0$. Since f is continuous and $f(t) = f(-t)$, both conclusions will follow if we show that $f(t) \leq \frac{1}{2}[f(t+s) + f(t-s)]$, whenever $t \pm s$ are in $[-1, 1]$.

For each t, let $\mathcal{M}_t$ be the mapping

$$\mathcal{M}_t(Y) = \frac{1}{2}(A^tYB^{-t} + A^{-t}YB^t).$$

For each Y, we have

$$|||Y||| = |||A^t(A^{-t}YB^{-t})B^t||| \leq \frac{1}{2}|||A^tYB^{-t} + A^{-t}YB^t|||$$

by Theorem IX.4.5. Thus $|||Y||| \leq |||\mathcal{M}_t(Y)|||$. From this it follows that

$$|||\mathcal{M}_t(AXB)||| \leq |||\mathcal{M}_s\mathcal{M}_t(AXB)|||, \quad \text{for all } s, t.$$

But,

$$\mathcal{M}_s\mathcal{M}_t = \frac{1}{2}(\mathcal{M}_{t+s} + \mathcal{M}_{t-s}).$$

So we have

$$|||\mathcal{M}_t(AXB)||| \leq \frac{1}{2}\{|||\mathcal{M}_{t+s}(AXB)||| + |||\mathcal{M}_{t-s}(AXB)|||\}.$$

Since $|||\mathcal{M}_t(AXB)||| = \frac{1}{2}f(t)$, this shows that

$$f(t) \leq \frac{1}{2}[f(t+s) + f(t-s)].$$

This proves the theorem. ■

Corollary IX.4.9 *Let A, B be positive matrices and X any matrix. Then, for each unitarily invariant norm, the function*

$$g(\nu) = |||A^\nu XB^{1-\nu} + A^{1-\nu}XB^\nu||| \tag{IX.28}$$

is convex on $[0, 1]$.

Proof. Replace A, B by $A^{1/2}$, $B^{1/2}$ in (IX.27). Then put $\nu = \frac{1+t}{2}$. ■

Corollary IX.4.10 *Let A, B be positive matrices and let X be any matrix. Then, for $0 \leq \nu \leq 1$ and for every unitarily invariant norm,*

$$|||A^\nu XB^{1-\nu} + A^{1-\nu}XB^\nu||| \leq |||AX + XB|||. \tag{IX.29}$$

Proof. Let $g(\nu)$ be the function defined in (IX.28). Note that $g(1) = g(0)$. So the assertion follows from the convexity of g. ■

IX.5 Schwarz Inequalities

In this section we shall prove some inequalities that can be considered to be matrix versions of the Cauchy-Schwarz inequality.

Some inequalities of this kind have already been proved in Chapter 4. Let A, B be any two matrices and r any positive real number. Then, we saw that

$$|||\, |A^*B|^r|||^2 \leq |||(AA^*)^r||| \;\; |||(BB^*)^r||| \tag{IX.30}$$

for every unitarily invariant norm. The choice $r = \frac{1}{2}$ gives the inequality

$$|||\, |A^*B|^{1/2}|||^2 \leq |||A||| \;\; |||B|||, \tag{IX.31}$$

while the choice $r = 1$ gives

$$|||A^*B|||^2 \leq |||AA^*||| \;\; |||BB^*|||. \tag{IX.32}$$

See Exercise IV.2.7 and Problem IV.5.7. It was noted there that the inequality (IX.32) is included in (IX.31).

We will now obtain more general versions of these in the same spirit as of Theorem IX.4.5. The generalisation of (IX.32) is proved easily and is given first, even though this is subsumed in the theorem that follows it.

Theorem IX.5.1 *Let A, B, X be any three matrices. Then, for every unitarily invariant norm,*

$$|||A^*XB|||^2 \leq |||AA^*X||| \;\; |||XBB^*|||. \tag{IX.33}$$

Proof. First assume that X is a positive matrix. Then

$$\begin{aligned} |||A^*XB|||^2 &= |||A^*X^{1/2}X^{1/2}B|||^2 = |||(X^{1/2}A)^*(X^{1/2}B)|||^2 \\ &\leq |||X^{1/2}AA^*X^{1/2}||| \;\; |||X^{1/2}BB^*X^{1/2}|||, \end{aligned}$$

using the inequality (IX.32). Now use Proposition IX.1.1 to conclude that

$$|||A^*XB|||^2 \leq |||AA^*X||| \;\; |||XBB^*|||.$$

This proves the theorem in this special case. Now let X be any matrix, and let $X = UP$ be its polar decomposition. Then, by unitary invariance,

$$\begin{aligned} |||A^*XB||| &= |||A^*UPB||| = |||U^*A^*UPB|||, \\ |||AA^*X||| &= |||AA^*UP||| = |||U^*AA^*UP|||, \\ |||XBB^*||| &= |||UPBB^*||| = |||PBB^*|||. \end{aligned}$$

So, the general theorem follows by applying the special case to the triple U^*AU, B, P. ■

The corresponding generalisation of the inequality (IX.30) is proved in the next theorem.

Theorem IX.5.2 *Let A, B, X be any three matrices. Then, for every positive real number r, and for every unitarily invariant norm, we have*

$$||| \, |A^*XB|^r |||^2 \leq ||| \, |AA^*X|^r ||| \;\; ||| \, |XBB^*|^r |||. \tag{IX.34}$$

Proof. Let $X = UP$ be the polar decomposition of X. Then $A^*XB = A^*UPB = (P^{1/2}U^*A)^*P^{1/2}B$. So, from the inequality (IX.30), we have

$$||| \, |A^*XB|^r |||^2 \leq |||(P^{1/2}U^*AA^*UP^{1/2})^r||| \;\; |||(P^{1/2}BB^*P^{1/2})^r|||. \tag{IX.35}$$

Now note that

$$\lambda^r(P^{1/2}U^*AA^*UP^{1/2}) = \lambda^r(AA^*UPU^*).$$

Using Theorem IX.2.9, we have

$$\lambda^r(AA^*UPU^*) \prec_w \lambda^{r/2}([AA^*]^2[UPU^*]^2).$$

But $(UPU^*)^2 = UP^2U^* = XX^*$. Hence,

$$\lambda^{r/2}([AA^*]^2[UPU^*]^2) = s^r(AA^*X).$$

Thus,

$$\lambda^r(P^{1/2}U^*AA^*UP^{1/2}) \prec_w s^r(AA^*X),$$

and hence

$$|||(P^{1/2}U^*AA^*UP^{1/2})^r||| \leq ||| \, |AA^*X|^r |||. \tag{IX.36}$$

In the same way, we have

$$\begin{aligned} \lambda^r(P^{1/2}BB^*P^{1/2}) &= \lambda^r(PBB^*) \prec_w \lambda^{r/2}(P^2[BB^*]^2) \\ &= s^r(PBB^*) = s^r(XBB^*). \end{aligned}$$

Hence

$$|||(P^{1/2}BB^*P^{1/2})^r||| \leq ||| \, |XBB^*|^r |||. \tag{IX.37}$$

Combining the inequalities (IX.35), (IX.36), and (IX.37) we get (IX.34). ■

The following corollary of Theorem IX.5.1 should be compared with (IX.29).

Corollary IX.5.3 *Let A, B be positive matrices and let X be any matrix. Then, for $0 \leq \nu \leq 1$, and for every unitarily invariant norm*

$$|||A^\nu XB^{1-\nu}||| \leq |||AX|||^\nu |||XB|||^{1-\nu}. \tag{IX.38}$$

Proof. For $\nu = 0, 1$, the inequality (IX.38) is a trivial statement. For $\nu = \frac{1}{2}$, it reduces to the inequality (IX.33). We will prove it, by induction, for all indices $\nu = k/2^n$, $k = 0, 1, \ldots, 2^n$. The general case then follows by continuity.

Let $\nu = \frac{2k+1}{2^n}$ be any dyadic rational. Then $\nu = \mu + \rho$, where $\mu = \frac{k}{2^{n-1}}$, $\rho = \frac{1}{2^n}$. Suppose that the inequality (IX.38) is valid for all dyadic rationals with denominator 2^{n-1}. Two such rationals are μ and $\lambda = \mu + 2\rho = \nu + \rho$. Then, using the inequality (IX.33) and this induction hypothesis, we have

$$\begin{aligned} |||A^\nu X B^{1-\nu}||| &= |||A^{\mu+\rho} X B^{1-\lambda+\rho}||| \\ &= |||A^\rho (A^\mu X B^{1-\lambda}) B^\rho||| \\ &\leq |||A^{2\rho} A^\mu X B^{1-\lambda}|||^{1/2} |||A^\mu X B^{1-\lambda} B^{2\rho}|||^{1/2} \\ &= |||A^\lambda X B^{1-\lambda}|||^{1/2} |||A^\mu X B^{1-\mu}|||^{1/2} \\ &\leq |||AX|||^{\lambda/2} |||XB|||^{(1-\lambda)/2} |||AX|||^{\mu/2} |||XB|||^{(1-\mu)/2} \\ &= |||AX|||^{(\lambda+\mu)/2} |||XB|||^{1-(\lambda+\mu)/2} \\ &= |||AX|||^\nu |||XB|||^{1-\nu}. \end{aligned}$$

This proves that the desired inequality holds for all dyadic rationals. ■

Corollary IX.5.4 *Let A, B be positive matrices and let X be any matrix. Then, for $0 \leq \nu \leq 1$, and for every unitarily invariant norm*

$$|||A^\nu X B^\nu||| \leq |||X|||^{1-\nu} |||AXB|||^\nu. \tag{IX.39}$$

Proof. Assume without loss of generality that A is invertible; the general case follows from this by continuity. We have, using (IX.38),

$$\begin{aligned} |||A^\nu X B^\nu||| &= |||(A^{-1})^{1-\nu} AXB^{1-(1-\nu)}||| \\ &\leq |||A^{-1}AX|||^{1-\nu} |||AXB|||^\nu \\ &= |||X|||^{1-\nu} |||AXB|||^\nu. \end{aligned}$$

■

Note that the inequality (IX.5) is a very special case of (IX.39).

Exercise IX.5.5 *Since $|||AA^*||| = |||A^*A|||$, the stars in the inequality (IX.32) could have been placed differently. Much less freedom is allowed for the generalisation (IX.33). Find a 2×2 example in which $|||A^*XB|||^2$ is larger than $|||A^*AX|||\ |||XBB^*|||$.*

Apart from norms, there are other interesting functions for which Schwarz-like inequalities can be obtained. This is done below. It is convenient to have a name for the class of functions we shall study.

Definition IX.5.6 *A continuous complex-valued function f on the space of matrices will be said to belong to the class $\mathcal{L}$ if it satisfies the following two conditions:*

(i) $f(B) \geq f(A) \geq 0$ if $B \geq A \geq 0$.

(ii) $|f(A^*B)|^2 \leq f(A^*A) f(B^*B)$ for all A, B.

We have seen above that every unitarily invariant norm is a function in the class $\mathcal{L}$. Other examples are given below.

Exercise IX.5.7 *(i) The functions trace and determinant are in $\mathcal{L}$.*

(ii) The function spectral radius is in $\mathcal{L}$.

(iii) If f is a function defined on matrices of order $\binom{n}{k}$ and is in $\mathcal{L}$, then the function $g(A) = f(\wedge^k A)$ defined on matrices of order n is also in $\mathcal{L}$.

(iv) The functions $\varphi_k(A) = \operatorname{tr} \wedge^k A$, $1 \leq k \leq n$, are in $\mathcal{L}$. (These are the coefficients in the characteristic polynomial of A.)

(v) If $s_j(A)$, $1 \leq j \leq n$, are the singular values of A, then for each $1 \leq k \leq n$ the function $f_k(A) = \prod_{j=1}^{k} s_j(A)$ is in $\mathcal{L}$.

(vi) If $\lambda_j(A)$ denote the eigenvalues of A arranged as $|\lambda_1(A)| \geq \cdots \geq |\lambda_n(A)|$, then for $1 \leq k \leq n$ the function $f_k(A) = \prod_{j=1}^{k} \lambda_j(A)$ is in $\mathcal{L}$.

Exercise IX.5.8 *Another class of functions $\mathcal{T}$ was introduced in IX.3.2. The two classes $\mathcal{T}$ and $\mathcal{L}$ have several elements in common. Find examples to show that neither of them is contained in the other.*

A different characterisation of the class $\mathcal{L}$ is obtained below. For this we need the following theorem, which is also useful in other contexts.

Theorem IX.5.9 *Let A, B be positive operators on $\mathcal{H}$, and let C be any operator on $\mathcal{H}$. Then the operator $\begin{pmatrix} A & C^* \\ C & B \end{pmatrix}$ on $\mathcal{H} \oplus \mathcal{H}$ is positive if and only if there exists a contraction K on $\mathcal{H}$ such that $C = B^{1/2} K A^{1/2}$.*

Proof. By Proposition I.3.5, K is a contraction if and only if $\begin{pmatrix} I & K^* \\ K & I \end{pmatrix}$ is positive. The positivity of this matrix implies the positivity of the matrix

$$\begin{pmatrix} A^{1/2} & 0 \\ 0 & B^{1/2} \end{pmatrix} \begin{pmatrix} I & K^* \\ K & I \end{pmatrix} \begin{pmatrix} A^{1/2} & 0 \\ 0 & B^{1/2} \end{pmatrix}$$

$$= \begin{pmatrix} A & A^{1/2} K^* B^{1/2} \\ B^{1/2} K A^{1/2} & B \end{pmatrix}.$$

(See Lemma V.1.5.) This proves one of the asserted implications. To prove the converse, first note that if A and B are invertible, then the argument can be reversed; and then note that the general case can be obtained from this by continuity. ∎

Theorem IX.5.10 *A (continuous) function f is in the class $\mathcal{L}$ if and only if it satisfies the following two conditions:*

(a) $f(A) \geq 0$ *for all* $A \geq 0$.

(b) $|f(C)|^2 \leq f(A)f(B)$ *for all* A, B, C *such that* $\begin{pmatrix} A & C^* \\ C & B \end{pmatrix}$ *is positive.*

Proof. If f satisfies condition (i) in Definition IX.5.6, then it certainly satisfies the condition (a) above. Further, if $\begin{pmatrix} A & C^* \\ C & B \end{pmatrix}$ is positive, then, by the preceding theorem, $C = B^{1/2}KA^{1/2}$, where K is a contraction. So, if f satisfies condition (ii) in IX.5.6, then

$$|f(C)|^2 = |f(B^{1/2}KA^{1/2})|^2 \leq f(B)f(A^{1/2}K^*KA^{1/2}).$$

Since $A^{1/2}K^*KA^{1/2} \leq A$, we also have $f(A^{1/2}K^*KA^{1/2}) \leq f(A)$ from the condition (i) in IX.5.6.

Now suppose f satisfies conditions (a) and (b). Let $B \geq A \geq 0$. Write

$$\begin{pmatrix} B & A \\ A & B \end{pmatrix} = \begin{pmatrix} B-A & 0 \\ 0 & B-A \end{pmatrix} + \begin{pmatrix} A & A \\ A & A \end{pmatrix}.$$

The first matrix in this sum is obviously positive; the second is also positive by Corollary I.3.3. Thus the sum is also positive. So it follows from (a) and (b) that $f(A) \leq f(B)$. Next note that we can write, for any A and B,

$$\begin{pmatrix} A^*A & A^*B \\ B^*A & B^*B \end{pmatrix} = \begin{pmatrix} A^* & 0 \\ B^* & 0 \end{pmatrix} \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}.$$

Since the two matrices on the right-hand side are adjoints of each other, their product is positive. Hence we have

$$|f(A^*B)|^2 \leq f(A^*A)f(B^*B).$$

This shows that f is in $\mathcal{L}$. ∎

This characterisation leads to an easy proof of the following theorem of E.H. Lieb.

Theorem IX.5.11 *Let $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$ be any matrices. Then, for every function f in the class $\mathcal{L}$,*

$$\left| f\left(\sum_{i=1}^m A_i^* B_i \right) \right|^2 \leq f\left(\sum_{i=1}^m A_i^* A_i \right) f\left(\sum_{i=1}^m B_i^* B_i \right). \qquad \text{(IX.40)}$$

$$\left| f\left(\sum_{i=1}^m A_i \right) \right|^2 \leq f\left(\sum_{i=1}^m |A_i| \right) f\left(\sum_{i=1}^m |A_i^*| \right). \qquad \text{(IX.41)}$$

Proof. For each $i = 1, \ldots, m$, the matrix $\begin{pmatrix} A_i^* A_i & A_i^* B_i \\ B_i^* A_i & B_i^* B_i \end{pmatrix}$ is positive, being the product of $\begin{pmatrix} A_i^* & 0 \\ B_i^* & 0 \end{pmatrix}$ and its adjoint. The sum of these matrices is, therefore, also positive. So the inequality (IX.40) follows from Theorem IX.5.10.

Each of the matrices $\begin{pmatrix} |A_i| & A_i^* \\ A_i & |A_i^*| \end{pmatrix}$ is also positive; see Corollary I.3.4. So the inequality (IX.41) follows by the same argument. ■

Exercise IX.5.12 *For any two matrices A, B, we have* $\operatorname{tr}(|A+B|) \le \operatorname{tr}(|A| + |B|)$. *Show by an example that this inequality is not true if* tr *is replaced by* det. *Show that we have*

$$[\det(|A+B|)]^2 \le \det(|A| + |B|) \det(|A^*| + |B^*|). \tag{IX.42}$$

A similar inequality holds for every function in $\mathcal{L}$.

IX.6 The Lieb Concavity Theorem

Let $f(A, B)$ be a real valued function of two matrix variables. Then, f is called **jointly concave**, if for all $0 \le \alpha \le 1$,

$$f(\alpha A_1 + (1-\alpha)A_2,\ \alpha B_1 + (1-\alpha)B_2) \ge \alpha f(A_1, B_1) + (1-\alpha) f(A_2, B_2)$$

for all A_1, A_2, B_1, B_2.

In this section we will prove the following theorem due to E.H. Lieb. The importance of the theorem, and its consequences, are explained later.

Theorem IX.6.1 *(Lieb) For each matrix X and each real number $0 \le t \le 1$, the function*

$$f(A, B) = \operatorname{tr} X^* A^t X B^{1-t}$$

is jointly concave on pairs of positive matrices.

Note that $f(A, B)$ is positive if A, B are positive.

To prove this theorem we need the following lemma.

Lemma IX.6.2 *Let $R_1, R_2, S_1, S_2, T_1, T_2$ be positive operators on a Hilbert space. Suppose R_1 commutes with R_2, S_1 commutes with S_2, and T_1 commutes with T_2, and*

$$R_1 \ge S_1 + T_1, \quad R_2 \ge S_2 + T_2. \tag{IX.43}$$

Then, for $0 \le t \le 1$,

$$R_1^t R_2^{1-t} \ge S_1^t S_2^{1-t} + T_1^t T_2^{1-t}. \tag{IX.44}$$

Proof. Let E be the set of all t in $[0,1]$ for which the inequality (IX.44) is true. It is clear that E is a closed set and it contains 0 and 1. We will first show that E contains the point $\frac{1}{2}$, then use this to show that E is a convex set. This would prove the lemma.

Let x, y be any two vectors. Then

$$\begin{aligned}
&|\langle x, (S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})y\rangle| \\
&\leq |\langle x, S_1^{1/2}S_2^{1/2}y\rangle| + |\langle x, T_1^{1/2}T_2^{1/2}y\rangle| \\
&\leq \|S_1^{1/2}x\|\,\|S_2^{1/2}y\| + \|T_1^{1/2}x\|\,\|T_2^{1/2}y\| \\
&\leq [\|S_1^{1/2}x\|^2 + \|T_1^{1/2}x\|^2]^{1/2}[\|S_2^{1/2}y\|^2 + \|T_2^{1/2}y\|^2]^{1/2}
\end{aligned}$$

by the Cauchy-Schwarz inequality. This last expression can be rewritten as

$$[\langle x, (S_1+T_1)x\rangle]^{1/2}[\langle y, (S_2+T_2)y\rangle]^{1/2}.$$

Hence, by the hypothesis, we have

$$|\langle x, (S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})y\rangle| \leq [\langle x, R_1x\rangle\langle y, R_2y\rangle]^{1/2}.$$

Using this, we see that, for all unit vectors u and v,

$$\begin{aligned}
&|\langle u, R_1^{-1/2}(S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})R_2^{-1/2}v\rangle| \\
&= |\langle R_1^{-1/2}u, (S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})R_2^{-1/2}v\rangle| \\
&\leq [\langle R_1^{-1/2}u,\; R_1^{1/2}u\rangle\langle R_2^{-1/2}v,\; R_2^{1/2}v\rangle]^{1/2} \\
&= 1.
\end{aligned}$$

This shows that

$$\|R_1^{-1/2}(S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})R_2^{-1/2}\| \leq 1.$$

Using Proposition IX.1.1, the commutativity of R_1 and R_2, and the inequality above, we see that

$$\|R_2^{-1/4}R_1^{-1/4}(S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})R_2^{-1/4}R_1^{-1/4}\| \leq 1.$$

This is equivalent to the operator inequality

$$R_2^{-1/4}R_1^{-1/4}(S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2})R_2^{-1/4}R_1^{-1/4} \leq I.$$

From this, it follows that

$$S_1^{1/2}S_2^{1/2} + T_1^{1/2}T_2^{1/2} \leq R_1^{1/2}R_2^{1/2}.$$

(See Lemma V.1.5.) This shows that the set E contains the point 1/2.

Now suppose that μ and ν are in E. Then

$$\begin{aligned} R_1^\mu R_2^{1-\mu} &\geq S_1^\mu S_2^{1-\mu} + T_1^\mu T_2^{1-\mu} \\ R_1^\nu R_2^{1-\nu} &\geq S_1^\nu S_2^{1-\nu} + T_1^\nu T_2^{1-\nu}. \end{aligned}$$

These two inequalities are exactly of the form (IX.43). Hence, using the special case of the lemma that has been proved above, we have,

$$\begin{aligned} &(R_1^\mu R_2^{1-\mu})^{1/2}(R_1^\nu R_2^{1-\nu})^{1/2} \\ &\geq (S_1^\mu S_2^{1-\mu})^{1/2}(S_1^\nu S_2^{1-\nu})^{1/2} + (T_1^\mu T_2^{1-\mu})^{1/2}(T_1^\nu T_2^{1-\nu})^{1/2}. \end{aligned}$$

Using the hypothesis about commutativity, we see from this inequality that $\frac{1}{2}(\mu+\nu)$ is in E. This shows that E is convex. ■

Proof of Theorem IX.6.1: Let A, B be positive operators on $\mathcal{H}$. Let $\mathcal{A}$ and $\mathcal{B}$ be the left and right multiplication operators on the space $\mathcal{L}(\mathcal{H})$ induced by A and B; i.e., $\mathcal{A}(X) = AX$ and $\mathcal{B}(X) = XB$. Using the results of Exercise I.4.4 and Problem VII.6.10, one sees that $\mathcal{A}$ and $\mathcal{B}$ are positive operators on (the Hilbert space) $\mathcal{L}(\mathcal{H})$.

Now, suppose A_1, A_2, B_1, B_2 are positive operators on $\mathcal{H}$. Let $A = A_1 + A_2$, $B = B_1 + B_2$. Let $\mathcal{A}_1$, $\mathcal{A}_2, \mathcal{A}$ denote the left multiplication operators on $\mathcal{L}(\mathcal{H})$ induced by A_1, A_2, and A, respectively, and $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}$ the right multiplication operators induced by B_1, B_2, and B, respectively. Then $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$, $\mathcal{B} = \mathcal{B}_1 + \mathcal{B}_2$. Hence, by Lemma IX.6.2.,

$$\mathcal{A}^t\mathcal{B}^{1-t} \geq \mathcal{A}_1^t\mathcal{B}_1^{1-t} + \mathcal{A}_2^t\mathcal{B}_2^{1-t},$$

for $0 \leq t \leq 1$. This is the same as saying that for every X in $\mathcal{L}(\mathcal{H})$

$$\langle X, A^tXB^{1-t}\rangle \geq \langle X, A_1^tXB_1^{1-t} + A_2^tXB_2^{1-t}\rangle,$$

or that

$$\operatorname{tr} X^*A^tXB^{1-t} \geq \operatorname{tr} X^*A_1^tXB_1^{1-t} + \operatorname{tr} X^*A_2^tXB_2^{1-t}.$$

From this, it follows that

$$f(\frac{A_1+A_2}{2}, \frac{B_1+B_2}{2}) \geq \frac{1}{2}f(A_1, B_1) + \frac{1}{2}f(A_2, B_2).$$

This shows that f is concave. ■

Another proof of this theorem is outlined in Problem IX.8.17.

Using the identification of $\mathcal{L}(\mathcal{H})$ with $\mathcal{H}\otimes\mathcal{H}$, Lieb's Theorem can be seen to be equivalent to the following theorem.

Theorem IX.6.3 *(T. Ando) For each $0 \leq t \leq 1$, the map $(A, B) \rightarrow A^t \otimes B^{1-t}$ is jointly concave on pairs of positive operators on $\mathcal{H}$.*

Exercise IX.6.4 *Let t_1, t_2 be two positive numbers such that $t_1 + t_2 \leq 1$. Show that the map $(A, B) \rightarrow A^{t_1} \otimes B^{t_2}$ is jointly concave on pairs of positive operators on $\mathcal{H}$. [Hint: The map $A \rightarrow A^s$ is monotone and concave on positive operators for $0 \leq s \leq 1$. See Chapter 5.]*

Lieb proved his theorem in connection with problems connected with entropy in quantum mechanics. This is explained below (with some simplifications).

The function $S(A) = -\text{tr}\, A \log A$, where A is a positive matrix, is called the **entropy** function. This is a concave function on the set of positive operators. In fact, we have seen that the function $f(t) = -t \log t$ is operator concave on $(0, \infty)$.

Let K be a given Hermitian operator. The **entropy** of A **relative** to K is defined as

$$S(A, K) = \frac{1}{2} \text{ tr } [A^{1/2}, K]^2,$$

where $[X, Y]$ stands for the Lie bracket (or the commutator) $XY - YX$. This concept was introduced by Wigner and Yanase, and extended by Dyson, who considered the functions

$$S_t(A, K) = \frac{1}{2} \text{ tr } ([A^t, K][A^{1-t}, K]), \tag{IX.45}$$

$0 < t < 1$.

The Wigner-Yanase-Dyson conjecture said that $S_t(A, K)$ is concave in A on the set of positive matrices. Lieb's Theorem implies that this is true. To see this note that

$$S_t(A, K) = \text{tr}(KA^tKA^{1-t} - K^2A). \tag{IX.46}$$

Since the function $g(A) = -\text{tr } K^2A$ is linear in A, it is also concave. Hence, concavity of $S_t(A, K)$ follows from that of tr KA^tKA^{1-t}. But that is a special case of Lieb's Theorem.

Given any operator X, define

$$I_t(A, X) = \text{tr}(X^*A^tXA^{1-t} - X^*XA), \tag{IX.47}$$

$0 \leq t < 1$. Note that $I_0(A, X) = 0$. When X is Hermitian, this reduces to the function S_t, defined earlier. Lieb's Theorem implies that $I_t(A, X)$ is a concave function of A. Hence, the function $I(A, X)$ defined as

$$I(A, X) = \left.\frac{d}{dt}\right|_{t=0} I_t(A, X) = \text{tr}(X^*(\log\ A)XA - X^*X(\log\ A)A) \tag{IX.48}$$

is also concave.

Let A, B be positive matrices. The **relative entropy** of A and B is defined as

$$S(A|B) = \text{tr}(A(\log\ A - \log\ B)). \tag{IX.49}$$

This notion was introduced by Umegaki, and generalised by Araki to the von Neumann algebra setting.

Theorem IX.6.5 *(Lindblad) The function $S(A|B)$ defined above is jointly convex in A, B.*

Proof. Consider the block-matrices

$$T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix}.$$

Then note that

$$S(A|B) = -I(T, X).$$

We noted earlier that $I(T, X)$ is concave in the argument T. ∎

Exercise IX.6.6 *(i) Show that for every pinching $\mathcal{C}$, $S(\mathcal{C}(A)|\mathcal{C}(B)) \leq S(A|B)$.*

(ii) Let τ be the normalised trace function on $n \times n$ matrices; i.e., $\tau(A) = \frac{1}{n} \operatorname{tr} A$. Show that for all positive matrices A, B

$$\tau(A)(\log \tau(A) - \log \tau(B)) \leq \tau(A(\log A - \log B)). \tag{IX.50}$$

This is called the **Peierls-Bogoliubov Inequality**. *(There are other inequalities that go by the same name.)*

IX.7 Operator Approximation

An operator approximation problem consists of finding, for a given operator A, the element nearest to it from a special class. Some problems of this type are studied in this section. In formulating and interpreting these results, it is helpful to have an analogy: if arbitrary operators are thought of as complex numbers, then Hermitian operators should be thought of as real numbers, unitary operators as complex numbers of modulus one and positive operators as positive real numbers. Of course, this analogy has its limitations, since multiplication of complex numbers is commutative and that of operators is not.

The first theorem below is easy to prove and sets the stage for later results.

Theorem IX.7.1 *Let A be any operator and let* $\operatorname{Re} A = \frac{1}{2}(A + A^*)$. *Then, for every Hermitian operator H and for every unitarily invariant norm,*

$$|||A - \operatorname{Re} A||| \leq |||A - H|||. \tag{IX.51}$$

Proof. Recall that $|||T||| = |||T^*|||$ for every T. Using this fact and the triangle inequality, we have

$$\begin{aligned} |||A - 1/2\,(A + A^*)||| &= 1/2\,|||A - A^*||| = 1/2\,|||A - H + H - A^*||| \\ &\leq 1/2\,(|||A - H||| + |||(A - H)^*|||) = |||A - H|||. \end{aligned}$$

This proves the theorem. ■

The inequality (IX.51) is sometimes stated in words as: in every unitarily invariant norm $\operatorname{Re} A$ is a **Hermitian approximant** to A.

The next theorem says that a **unitary approximant** to A is any unitary that occurs in its polar decomposition.

Theorem IX.7.2 *If $A = UP$, where U is unitary and P positive, then*

$$|||A - U||| \leq |||A - W||| \leq |||A + U||| \tag{IX.52}$$

for every unitary W and for every unitarily invariant norm.

Proof. By unitary invariance, the inequality (IX.52) is equivalent to

$$|||P - I||| \leq |||P - U^*W||| \leq |||P + I|||.$$

So the assertion of the theorem is equivalent to the following: for every positive operator P and unitary operator V,

$$|||P - I||| \leq |||P - V||| \leq |||P + I|||. \tag{IX.53}$$

This will be proved using the spectral perturbation inequality (IV.62). Let

$$\tilde{P} = \begin{pmatrix} 0 & P \\ P & 0 \end{pmatrix}, \quad \tilde{V} = \begin{pmatrix} 0 & V \\ V^* & 0 \end{pmatrix}.$$

Then $\tilde{P}$ and $\tilde{V}$ are Hermitian. The eigenvalues of $\tilde{P}$ are the singular values of P together with their negatives. (See Exercise II.1.15.) The same is true for $\tilde{V}$, which means that it has eigenvalues 1 and -1, each with multiplicity n. We thus have

$$\operatorname{Eig}^{\downarrow}(\tilde{P}) - \operatorname{Eig}^{\downarrow}(\tilde{V}) = [\operatorname{Eig}^{\downarrow}(P) - I] \oplus [-\operatorname{Eig}^{\uparrow}(P) + I],$$

$$\operatorname{Eig}^{\downarrow}(\tilde{P}) - \operatorname{Eig}^{\uparrow}(\tilde{V}) = [\operatorname{Eig}^{\downarrow}(P) + I] \oplus [-\operatorname{Eig}^{\uparrow}(P) - I].$$

So, from (IV.62), we have

$$\begin{aligned} |||[\operatorname{Eig}^{\downarrow}(P) - I] \oplus [\operatorname{Eig}^{\downarrow}(P) - I]||| &\leq |||(P - V) \oplus (P - V)^*||| \\ &\leq |||[\operatorname{Eig}^{\downarrow}(P) + I] \oplus [\operatorname{Eig}^{\downarrow}(P) + I]|||. \end{aligned}$$

This is equivalent to the pair of inequalities (IX.53). ■

The two approximation problems solved above are subsumed in a more general question. Let Φ be a closed subset of the complex plane, and let $\mathbf{N}(\mathbf{\Phi})$ be the collection of all normal operators whose spectrum is contained in Φ. Given any operator A, what operator in $\mathbf{N}(\mathbf{\Phi})$ is closest to A? The two theorems proved above answer this when Φ is the real line or the unit

circle. When Φ is the whole plane or the positive half-line, the problem becomes much harder, and the full solution is not known. Note that in the first case ($\Phi = \mathbb{C}$) we are asking for a normal approximant to A, and in the second case ($\Phi = \mathbb{R}_+$) for a positive approximant to A. Some results on this problem, which are easy to describe and also are directly related to other parts of this book, are given below.

We have already come across a special case of this problem in Chapter 6. Let F be a retraction of the plane onto the subset Φ; i.e., F is a map of $\mathbb{C}$ onto Φ such that $|z - F(z)| \le |z - w|$ for all $z \in \mathbb{C}$ and $w \in \Phi$. Such a map always exists; it is unique if (and only if) Φ is convex. We have the following theorem.

Theorem IX.7.3 *Let F be a retraction of the plane onto the closed set Φ. Suppose Φ is convex. Then, for every normal operator A, we have*

$$|||A - F(A)||| \le |||A - N||| \tag{IX.54}$$

for all $N \in \mathbf{N}(\mathbf{\Phi})$ and for all unitarily invariant norms. If the set Φ is not convex, the inequality (IX.54) may not be true for all unitarily invariant norms, but is still true for all Q-norms. (See Theorem VI.6.2 and Problem VI.8.13.)

Exercise IX.7.4 *Let A be a Hermitian operator, and let $A = A^+ - A^-$ be its Jordan decomposition. (Both A^+ and A^- are positive operators.) Use the above theorem to show that, if P is any positive operator, then*

$$|||A - A^+||| \le |||A - P||| \tag{IX.55}$$

for every unitarily invariant norm. If A is normal, then for every positive operator P

$$|||A - (\operatorname{Re} A)^+||| \le |||A - P|||. \tag{IX.56}$$

Theorem IX.7.5 *Let A be any operator. Then for every positive operator P*

$$\|A - (\operatorname{Re} A)^+\|_2 \le \|A - P\|_2. \tag{IX.57}$$

Proof. Recall that $\|A\|_2^2 = \|\operatorname{Re} A\|_2^2 + \|\operatorname{Im} A\|_2^2$. Hence,

$$\|A - (\operatorname{Re} A)^+\|_2^2 = \|\operatorname{Re} A - (\operatorname{Re} A)^+\|_2^2 + \|\operatorname{Im} A\|_2^2.$$

From (IX.55), we see that $\|\operatorname{Re} A - (\operatorname{Re} A)^+\|_2^2$ is bounded by $\|\operatorname{Re} A - P\|_2^2$. This leads to the inequality (IX.57). ∎

The problem of finding positive approximants to an arbitrary operator A is much more complex for other norms. See the Notes at the end of the chapter.

For normal approximants, we have a solution in all unitarily invariant norms only in the 2×2 case.

Theorem IX.7.6 *Let A be an upper triangular matrix*

$$A = \begin{pmatrix} \lambda_1 & b \\ 0 & \lambda_2 \end{pmatrix}, \quad b \geq 0. \tag{IX.58}$$

Let $\theta = \arg(\lambda_1 - \lambda_2)$, and let

$$N_0 = \begin{pmatrix} \lambda_1 & \frac{1}{2}b \\ \frac{1}{2}e^{2i\theta}b & \lambda_2 \end{pmatrix}. \tag{IX.59}$$

Then N_0 is normal, and for any normal matrix N we have

$$|||A - N_0||| \leq |||A - N||| \tag{IX.60}$$

for every unitarily invariant norm.

Proof. It is easy to check that N_0 is normal. Let $N = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}$ be any normal matrix. We must have $|x_2| = |x_3|$ in that case.

Now note that, if $T = \begin{pmatrix} t_1 & t_2 \\ t_3 & t_4 \end{pmatrix}$ is any matrix, we can write its off-diagonal part $\begin{pmatrix} 0 & t_2 \\ t_3 & 0 \end{pmatrix}$ as $\frac{1}{2}(T - UTU^*)$, where U is the diagonal matrix with diagonal entries 1 and -1. Hence, for every unitarily invariant norm, we have

$$|||T||| \geq \left|\left|\left| \begin{pmatrix} 0 & t_2 \\ t_3 & 0 \end{pmatrix} \right|\right|\right| = \left|\left|\left| \begin{pmatrix} t_2 & 0 \\ 0 & t_3 \end{pmatrix} \right|\right|\right|.$$

Using this, we see that $|||A - N||| \geq |||\operatorname{diag}(b - x_2, -x_3)|||$. But,

$$b \leq |b - x_2| + |x_2| = |b - x_2| + |x_3| \leq 2 \max(|b - x_2|, |x_3|).$$

Thus the vector $\frac{1}{2}(b, b)$ is weakly majorised by the vector $(|b - x_2|, |x_3|)$, which, in turn, is weakly majorised by the vector $(s_1(A - N), s_2(A - N))$ as seen above. Since $A - N_0$ has singular values $(\frac{1}{2}b, \frac{1}{2}b)$, this proves the inequality (IX.60). ■

Since every 2×2 matrix is unitarily equivalent to an upper triangular matrix of the form (IX.58), this theorem tells us how to find a normal matrix closest to it.

Exercise IX.7.7 *The measure of nonnormality of a matrix, with respect to any norm, was defined in Problems VIII.6.4 and VIII.6.5. Theorem IX.7.6, on the other hand, gives for 2×2 matrices a formula for the distance to the set of all normal matrices. What is the relation between these two numbers for a given unitarily invariant norm?*

IX.8 Problems

Problem IX.8.1. Let A, B be positive matrices, and let m, k be positive integers with $m \geq k$. Use the inequality (IX.13) to show that

$$\operatorname{tr}(A^k B^k)^m \leq \operatorname{tr}(A^m B^m)^k. \tag{IX.61}$$

The special case,

$$\operatorname{tr}(AB)^m \leq \operatorname{tr} A^m B^m, \tag{IX.62}$$

is called the **Lieb-Thirring inequality.**

Problem IX.8.2. Let A, B be Hermitian matrices. Show that for every positive integer m

(i) $|\operatorname{tr}(AB)^{2m}| \leq \operatorname{tr} A^{2m} B^{2m}$,

(ii) $|\operatorname{tr}(A^m B^m)^2| \leq \operatorname{tr} A^{2m} B^{2m}$,

(iii) $|\operatorname{tr}(AB)^{4m}| \leq \operatorname{tr}(A^{2m} B^{2m})^2$.

(Hint: By the Weyl Majorant Theorem $|\operatorname{tr} X^m| \leq \operatorname{tr}|X|^m$, for every matrix X.) Note that if

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix},$$

then $|\operatorname{tr}(AB)^3| = 5, |\operatorname{tr} A^3 B^3| = 4$, $\operatorname{tr}(AB)^6 = 9$, and $\operatorname{tr}(A^3 B^3)^2 = 0$. This shows the failure of possible extensions of the inequalities (i) and (iii) above.

Problem IX.8.3. Are there any natural generalisations of the above inequalities when three matrices A, B, C are involved? Take, for instance, the inequality (IX.62). A product of three positive matrices need not have positive eigenvalues. One still might wonder whether $|\operatorname{tr}(ABC)^2| \leq |\operatorname{tr} A^2 B^2 C^2|$. Construct an example to show that this need not be true.

Problem IX.8.4. A possible generalisation of the Golden-Thompson inequality (IX.19) would have been $\operatorname{tr}(e^{A+B+C}) \leq |\operatorname{tr}(e^A e^B e^C)|$ for any three Hermitian matrices A, B, C. This is false. To see this, let S_1, S_2, S_3 be the Pauli spin matrices

$$S_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad S_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

If a_1, a_2, a_3 are any real numbers and $a = (a_1^2 + a_2^2 + a_3^2)^{1/2}$, show that

$$\exp(\Sigma a_j S_j) = (\cosh\, a)I + \frac{\sinh\, a}{a}\, \Sigma a_j S_j.$$

Let

$$A = tS_1, \;\; B = tS_2, \;\; C = t(S_3 - S_2 - S_1).$$

Show that

$$\begin{aligned} \operatorname{tr}(e^{A+B+C}) &= 2 \cosh t, \\ |\operatorname{tr}(e^A e^B e^C)| &= 2 \cosh t[1 - \frac{t^4}{12} + O(t^6)]. \end{aligned}$$

For small t, the first quantity is bigger than the second.

Problem IX.8.5. Show that the Lie Product Formula has a generalisation: for any k matrices $A_1, A_2, \ldots, A_k$,

$$\lim_{m\to\infty} \left(\exp \frac{A_1}{m} \exp \frac{A_2}{m} \cdots \exp \frac{A_k}{m} \right)^m = \exp(A_1 + A_2 + \cdots + A_k).$$

Problem IX.8.6. Show that for any two matrices A, B we have

$$|||A^*B + B^*A||| \leq |||AA^* + BB^*|||$$

and

$$|||A^*B + B^*A||| \leq |||A^*A + B^*B|||$$

for every unitarily invariant norm.

Problem IX.8.7. Let X, Y be positive. Show that for every unitarily invariant norm

$$|||X - Y||| \leq \left|\left|\left| \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} \right|\right|\right|.$$

From this, it follows that, for every A,

$$\|A^*A - AA^*\| \leq \|A\|^2,$$

and

$$\|A^*A - AA^*\|_p \leq 2^{1/p} \|A\|_{2p}^2, \quad 1 \leq p < \infty.$$

Problem IX.8.8. Let A, B be positive matrices and let X be any matrix. Show that for all unitarily invariant norms, and for $0 \leq \nu \leq 1$,

$$|||A^\nu X B^{1-\nu} - A^{1-\nu} X B^\nu||| \leq |2\nu - 1| \; |||AX - XB|||.$$

Problem IX.8.9. Let A, B be positive operators and let T be any operator such that $\|T^*x\| \leq \|Ax\|$ and $\|Tx\| \leq \|Bx\|$ for all x. Show that, for all x, y and for $0 \leq \nu \leq 1$,

$$|\langle x, Ty \rangle| \leq \|A^{1-\nu}x\| \; \|B^\nu y\|.$$

[Hint: From the hypotheses, it follows that $A^{-1}T$ and TB^{-1} are contractions. The inequality (IX.38) then implies that $(A^{-1})^{1-\nu}T(B^{-1})^{\nu}$ is a contraction.]

Problem IX.8.10. Use the result of the above problem to prove the following. For all operators T, vectors x, y, and for $0 \leq \nu \leq 1$,

$$|\langle x, Ty\rangle|^2 \leq \langle x, |T^*|^{2(1-\nu)}x\rangle \ \langle y, |T|^{2\nu}y\rangle.$$

This inequality is called the **Mixed Schwarz Inequality**.

Problem IX.8.11. Show that if A, B are positive matrices, then we have

$$\det(I + A + B) \leq \det(I + A)\det(I + B).$$

Then use this and Theorem IX.5.11 to show that, for any two matrices A, B,

$$|\det(I + A + B)| \leq \det(I + |A|)\det(I + |B|).$$

(See Problem IV.5.9 for another proof of this.)

Problem IX.8.12. Show that for all positive matrices A, B

$$\text{tr}(A(\log\ A - \log\ B)) \geq \text{tr}(A - B). \tag{IX.63}$$

The example $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 2 \end{pmatrix}$ shows that we may not have the operator inequality $A(\log A - \log B) \geq (A - B)$.

Problem IX.8.13. Let f be a convex function on an interval I. Let A, B be two Hermitian matrices whose spectra are contained in I. Show that

$$\text{tr}[f(A) - f(B)] \geq \text{tr}[(A - B)f'(B)]. \tag{IX.64}$$

The special choice $f(t) = t \log t$ gives the inequality (IX.63).

Problem IX.8.14. Let A be a Hermitian matrix and f any convex function. Then for every unit vector x

$$f(\langle x, Ax\rangle) \leq \langle x, f(A)x\rangle.$$

This implies that, for any orthonormal basis $x_1, \ldots, x_n$,

$$\sum_{j=1}^{n} f(\langle x_j, Ax_j\rangle) \leq \text{tr}\ f(A).$$

The name Peierls-Bogoliubov inequality is sometimes used for this inequality, or for its special cases $f(t) = e^t$, $f(t) = e^{-t}$, etc.

Problem IX.8.15. The concavity assertion in Exercise IX.6.4. can be generalised to several variables. Let t_1, t_2, t_3 be positive numbers such that $t_1 + t_2 + t_3 \le 1$. Let A_1, A_2, A_3 be positive operators. Note that

$$A_1^{t_1} \otimes A_2^{t_2} \otimes A_3^{t_3} = (A_1^{t_1} \otimes A_2^{t_2} \otimes I)(I \otimes I \otimes A_3^{t_3}).$$

Use the concavity of the first factor above (which has been proved in Exercise IX.6.4) and the integral representation (V.4) for the second factor to prove that the map $(A_1, A_2, A_3) \to A_1^{t_1} \otimes A_2^{t_2} \otimes A_3^{t_3}$ is jointly concave on triples of positive operators. More generally, prove that for positive numbers $t_1, \ldots, t_k$ with $t_1 + \cdots + t_k \le 1$, the map that takes a k-tuple of positive operators $(A_1, \ldots, A_k)$ to the operator $A_1^{t_1} \otimes \cdots \otimes A_k^{t_k}$ is jointly concave.

Problem IX.8.16. A special consequence of the above is that the map $A \to \otimes^k A^{1/k}$ is concave on positive operators for all $k = 1, 2, \ldots$ Use this to prove the following inequalities for $n \times n$ positive matrices A, B:

$$\begin{array}{lll}
\text{(i)} & \otimes^k (A+B)^{1/k} & \ge \otimes^k A^{1/k} + \otimes^k B^{1/k}, \\
\text{(ii)} & \wedge^k (A+B)^{1/k} & \ge \wedge^k A^{1/k} + \wedge^k B^{1/k}, \\
\text{(iii)} & \vee^k (A+B)^{1/k} & \ge \vee^k A^{1/k} + \vee^k B^{1/k}, \\
\text{(iv)} & \det(A+B)^{1/n} & \ge \det A^{1/n} + \det B^{1/n}, \\
\text{(v)} & \operatorname{per}(A+B)^{1/n} & \ge \operatorname{per} A^{1/n} + \operatorname{per} B^{1/n}, \\
\text{(vi)} & c_k((A+B)^{1/k}) & \ge c_k(A^{1/k}) + c_k(B^{1/k}),
\end{array}$$

where $c_k(A) = \operatorname{tr} \wedge^k (A)$ for $1 \le k \le n$.

The inequality (iv) above is called the Minkowski Determinant Theorem and has been proved earlier (Corollary II.3.21).

Problem IX.8.17. Outlined below is another proof of the Lieb Concavity Theorem which uses results on operator concave functions proved in Chapter 5.

(i) Consider the space $\mathcal{L}(\mathcal{H}) \oplus \mathcal{L}(\mathcal{H})$ with the inner product

$$\langle (R_1, R_2),\ (S_1, S_2) \rangle = \operatorname{tr}(R_1^* S_1 + R_2^* S_2).$$

(ii) Let A_1, A_2 be invertible positive operators on $\mathcal{H}$ and let $A = 1/2\ (A_1 + A_2)$. Let

$$\begin{aligned}
\Delta(R) &= ARA^{-1}, \\
\Delta_{12}(R, S) &= (A_1 R A_1^{-1},\ A_2 R A_2^{-1}).
\end{aligned}$$

Then Δ is a positive operator on the Hilbert space $\mathcal{L}(\mathcal{H})$ and Δ_{12} is a positive operator on the Hilbert space $\mathcal{L}(\mathcal{H}) \oplus \mathcal{L}(\mathcal{H})$

(iii) Note that for any X in $\mathcal{L}(\mathcal{H})$

$$\operatorname{tr} X^* A^t X A^{1-t} = \langle X A^{1/2},\ \Delta^t (X A^{1/2}) \rangle$$

and

$$\begin{aligned} &\operatorname{tr}(X^* A_1^t X A_1^{1-t} + X^* A_2^t X A_2^{1-t}) \\ &= \langle (XA_1^{1/2}, XA_2^{1/2}),\ \Delta_{12}^t (XA_1^{1/2}, XA_2^{1/2}) \rangle. \end{aligned}$$

(iv) Let V be the map from $\mathcal{L}(\mathcal{H})$ into $\mathcal{L}(\mathcal{H}) \oplus \mathcal{L}(\mathcal{H})$ defined as

$$V(XA^{1/2}) = \frac{1}{\sqrt{2}}(XA_1^{1/2},\ XA_2^{1/2}).$$

Show that V is an isometry. Show that

$$\Delta = V^* \Delta_{12} V.$$

(v) Since the function $f(t)$ on $[0, \infty)$ is operator concave for $0 < t < 1$, using Exercise V.2.4, we obtain

$$V^* \Delta_{12}^t V \leq (V^* \Delta_{12} V)^t = \Delta^t.$$

(vi) This shows that

$$\operatorname{tr} X^* A^t X A^{1-t} \geq 1/2 \operatorname{tr}(X^* A_1^t X A_1^{1-t} + X^* A_2^t X A_2^{1-t})$$

when A_1 and A_2 are invertible. By continuity, this is true for all positive operators A_1 and A_2. In other words, for all $0 < t < 1$, the function

$$f(A) = \operatorname{tr} X^* A^t X A^{1-t}$$

is concave.

(vii) Use 2×2 operator matrices $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ X & 0 \end{pmatrix}$ to complete the proof of Lieb's concavity theorem.

Problem IX.8.18. Theorem IX.7.1 can be generalised as follows. Let φ be a mapping of the space of $n \times n$ matrices into itself that satisfies three conditions:

(i) φ^2 is the identity map; i.e., $\varphi(\varphi(A)) = A$ for all A.

(ii) φ is real linear; i.e., $\varphi(\alpha A + \beta B) = \alpha\varphi(A) + \beta\varphi(B)$ for all A, B and all real α, β.

(iii) A and $\varphi(A)$ have the same singular values for all A.

Then the set $I(\varphi) = \{A : \varphi(A) = A\}$ is a real linear subspace of the space of matrices. For each A, the matrix $\frac{1}{2}(A + \varphi(A))$ is in $I(\varphi)$, and for all unitarily invariant norms $|||A - \frac{1}{2}(A + \varphi(A))||| \leq |||A - B|||$ for all $B \in I(\varphi)$.

Examples of such maps are $\varphi(A) = \pm A^*$, $\varphi(A) = \pm A^T$, and $\varphi(A) = \pm \bar{A}$, where A^T denotes the transpose of A and $\bar{A}$ denotes the matrix obtained by taking the complex conjugate of each entry of A.

Problem IX.8.19. The **Cayley transform** of a Hermitian matrix A is the unitary matrix $C(A)$ defined as

$$C(A) = (A - iI)(A + iI)^{-1}.$$

If A, B are two Hermitian matrices, we have

$$\frac{1}{2i}[C(A) - C(B)] = (B + iI)^{-1}(A - B)(A + iI)^{-1}.$$

Use this to show that for all j,

$$1/2 \; s_j(C(A) - C(B)) \leq s_j(A - B).$$

[Note that $\|(A+iI)^{-1}\| \leq 1$ and $\|(B+iI)^{-1}\| \leq 1$.] In particular, this gives

$$1/2 \; |||C(A) - C(B)||| \leq |||A - B|||$$

for every unitarily invariant norm.

Problem IX.8.20. A 2×2 block matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, in which the four matrices A_{ij} are normal and commute with each other, is called **binormal**. Show that such a matrix is unitarily equivalent to a matrix $\tilde{A} = \begin{pmatrix} \Lambda_1 & B \\ 0 & \Lambda_2 \end{pmatrix}$, in which Λ_1, Λ_2, B are diagonal matrices and B is positive. Let

$$N_0 = \begin{pmatrix} \Lambda_1 & \frac{1}{2}B \\ \frac{1}{2}U^2 B & \Lambda_2 \end{pmatrix},$$

where U is the unitary operator such that $\Lambda_1 - \Lambda_2 = U|\Lambda_1 - \Lambda_2|$. Show that in every unitarily invariant norm we have

$$|||\tilde{A} - N_0||| \leq |||\tilde{A} - N|||$$

for all $2n \times 2n$ normal matrices N.

Problem IX.8.21. An alternate proof of the inequality (IX.55) is outlined below. Choose an orthonormal basis in which A is diagonal and $A = \begin{pmatrix} A^+ & 0 \\ 0 & -A^- \end{pmatrix}$. In this basis let P have the block decomposition $P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$. By the pinching inequality,

$$|||A - P||| \geq \left|\left|\left| \begin{pmatrix} A^+ - P_{11} & 0 \\ 0 & -A^- - P_{22} \end{pmatrix} \right|\right|\right|.$$

Since both A^- and P_{22} are positive, $|||A^-||| \leq |||A^- + P_{22}|||$. Use this to prove the inequality (IX.55).

This argument can be modified to give another proof of (IX.56) also. For this we need the following fact. Let T and S be operators such that $0 \leq \operatorname{Re} T \leq \operatorname{Re} S$, and $\operatorname{Im} T = \operatorname{Im} S$. If T is normal, then $|||T||| \leq |||S|||$, for every unitarily invariant norm. Prove this using the Fan Dominance Theorem (Theorem IV.2.2) and the result of Problem III.6.6.

IX.9 Notes and References

Matrix inequalities of the kind studied in this chapter can be found in several books. The reader should see, particularly, the books by Horn and Johnson, Marshall and Olkin, Marcus and Minc, mentioned in Chapters 1 and 2; and, in addition, M.L. Mehta, *Matrix Theory*, second edition, Hindustan Publishing Co., 1989, and B. Simon, *Trace Ideals and Their Applications*, Cambridge University Press, 1979.

Proposition IX.1.1 is proved in B. Simon, *Trace Ideals*, p. 95. Proposition IX.1.2, for the operator norm, is proved in A. McIntosh, *Heinz inequalities and perturbation of spectral families*, Macquarie Math. Reports, 1979; and for all unitarily invariant norms in F. Kittaneh, *A note on the arithmetic-geometric mean inequality for matrices*, Linear Algebra Appl., 171 (1992) 1-8. The Lie product formula has been extended to semigroups of operators in Banach spaces by H.F. Trotter.

Our treatment of the material between Theorem IX.2.1 and Exercise IX.2.8 is based on T. Furuta, *Norm inequalities equivalent to Löwner-Heinz theorem*, Reviews in Math. Phys., 1(1989) 135-137. The inequality (IX.5) can also be found in H.O. Cordes, *Spectral Theory of Linear Differential Operators and Comparison Algebras*, Cambridge University Press, 1987. Theorem IX.2.9 is taken from B. Wang and M. Gong, *Some eigenvalue inequalities for positive semidefinite matrix power products*, Linear Algebra Appl., 184(1993) 249-260. The inequality (IX.13) was proved by H. Araki, *On an inequality of Lieb and Thirring*, Letters in Math. Phys., 19(1990) 167-170. Theorem IX.2.10 is a rephrasing of some other results proved in this paper.

The Golden-Thompson inequality is important in statistical mechanics. See S. Golden, *Lower bounds for the Helmholtz function*, Phys. Rev. B, 137 (1965) 1127-1128, and C.J. Thompson, *Inequality with applications in statistical mechanics*, J. Math. Phys., 6(1965) 1812-1813. It was generalised by A. Lenard, *Generalization of the Golden-Thompson inequality*, Indiana Univ. Math. J. 21(1971) 457-468, and further by C.J. Thompson, *Inequalities and partial orders on matrix spaces*, Indiana Univ. Math. J. 21 (1971) 469-480. These ideas have been developed further in the much more general setting of Lie groups by B. Kostant, *On convexity, the Weyl group and the Iwasawa decomposition*, Ann. Sci. E.N.S., 6(1973) 413-455, and subsequently by others. Inequalities complementary to the Golden-Thompson inequality and its stronger version in Exercise IX.3.8 have been proved by F. Hiai and D. Petz, *The Golden-Thompson trace inequality is complemented*, Linear Algebra Appl., 181(1993) 153-185, and by T. Ando and F. Hiai, *Log majorization and complementary Golden-Thompson type inequalities*, Linear Algebra Appl., 197/198(1994) 113-131. Theorem IX.3.1 is a rephrasing of some results in J.E. Cohen, *Inequalities for matrix exponentials*, Linear Algebra Appl., 111(1988) 25-28. Further results on such inequalities may be found in J.E. Cohen, S. Friedland, T. Kato, and F.P.

Kelly, *Eigenvalue inequalities for products of matrix exponentials*, Linear Algebra Appl., 45(1982) 55-95, in D. Petz, *A survey of certain trace inequalities*, Banach Centre Publications 30(1994) 287-298, and in Chapter 6 of Horn and Johnson, *Topics in Matrix Analysis.*

Theorem IX.4.2 was proved in R. Bhatia and F. Kittaneh, *On the singular values of a product of operators*, SIAM J. Matrix Analysis, 11(1990) 272-277. The generalisation given in Theorem IX.4.5 is due to R. Bhatia and C. Davis, *More matrix forms of the arithmetic-geometric mean inequality*, SIAM J. Matrix Analysis, 14(1993) 132-136. Many of the other results in Section IX.4 are from these two papers. The proof outlined in Exercise IX.4.6 is due to F. Kittaneh, *A note on the arithmetic-geometric mean inequality for matrices*, Linear Algbera Appl., 171(1992) 1-8. A generalisation of the inequality (IX.21) has been proved by T. Ando, *Matrix Young inequalities*, Operator Theory: Advances and Applications, 75(1995) 33-38. If $p,q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then the operator inequality $|AB^*| \leq U(\frac{1}{p}|A|^p + \frac{1}{q}|B|^q)U^*$ is valid for some unitary U.

Theorems IX.5.1 and IX.5.2 were proved in R. Bhatia and C. Davis, *A Cauchy-Schwarz inequality for operators with applications*, Linear Algebra Appl., 223(1995) 119-129. For the case of the operator norm, the inequality (IX.38) is due to E. Heinz, as are the inequality (IX.29) and the one in Problem IX.8.8. See E. Heinz, *Beiträge zur Störungstheorie der Spektralzerlegung*, Math. Ann., 123(1951) 415-438. Our approach to these inequalities follows the one in the paper by A. McIntosh cited above. The inequality in Problem IX.8.9 is also due to E. Heinz. The Mixed Schwarz inequality in Problem IX.8.10 was proved by T. Kato, *Notes on some inequalities for linear operators*, Math. Ann., 125(1952) 208-212. (The papers by Heinz, Kato, and McIntosh do much of this for unbounded operators in infinite-dimensional spaces.) The class $\mathcal{L}$ in Definition IX.5.6 was introduced by E.H. Lieb, *Inequalities for some operator and matrix functions*, Advances in Math., 20(1976) 174-178. Theorem IX.5.11 was proved in this paper. These functions are also studied in R. Merris and J.A. Dias da Silva, *Generalized Schur functions*, J. Algebra, 35(1975) 442-448. B. Simon (*Trace Ideals*, p. 99) calls them Liebian functions. The characterisation in Theorem IX.5.10 has not appeared before; it simplifies the proof of Theorem IX.5.11 considerably.

The Lieb Concavity Theorem was proved by E.H. Lieb, *Convex trace functions and the Wigner-Yanase-Dyson conjecture*, Advances in Math., 11(1973) 267-288. The proof given here is taken from B. Simon, *Trace Ideals.* T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26(1979) 203-241, takes a different approach. Using the concept of operator means, he first proves Theorem IX.6.3 (and its generalisation in Problem IX.8.15) and then deduces Lieb's Theorem from it. The proof given in Problem IX.8.17 is taken from D. Petz, *Quasi-entropies for finite quantum systems*, Rep.

Math. Phys., 21(1986) 57-65. Theorem IX.6.5 was proved by G. Lindblad, *Entropy, information and quantum measurements*, Commun. Math. Phys., 33(1973) 305-322. Our proof is taken from A. Connes and E. Stϕrmer, *Entropy for automorphisms of II_1 von Neumann algebras*, Acta Math., 134(1975) 289-306. The reader would have guessed from the titles of these papers that these inequalities are useful in physics. The book *Quantum Entropy and Its Use* by M. Ohya and D. Petz, Springer-Verlag, 1993, contains a very detailed study of such inequalities. Another pertinent reference is D. Ruelle, *Statistical Mechanics*, Benjamin, 1969. The inequalities in Problem IX.8.16 are taken from T. Ando, *Inequalities for permanents*, Hokkaido Math. J., 10(1981) 18-36, and R. Bhatia and C. Davis, *Concavity of certain functions of matrices*, Linear and Multilinear Algebra, 17 (1985) 155-164.

Theorems IX.7.1 and IX.7.2 were proved in K. Fan and A.J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6(1955) 111-116. The inequalities in Problem IX.8.19 were also proved in this paper. The result in Problem IX.8.18 is due to C.-K. Li and N.-K. Tsing, *On the unitarily invariant norms and some related results*, Linear and Multilinear Algebra, 20 (1987) 107-119. Two papers by P.R. Halmos, *Positive approximants of operators*, Indiana Univ. Math. J. 21(1972) 951-960, and *Spectral approximants of normal operators*, Proc. Edinburgh Math. Soc., 19 (1974) 51-58, made the problem of operator approximation popular among operator theorists. The results in Theorem IX.7.3, Exercise IX.7.4, and in Problem IX.8.21, were proved in these papers for the special case of the operator norm (but more generally for Hilbert space operators). The first paper of Halmos also tackles the problem of finding a positive approximant to an arbitrary operator, in the operator norm. The solution is different from the one for the Hilbert-Schmidt norm given in Theorem IX.7.5, and the problem is much more complicated. The problem of finding the closest normal matrix has been solved completely only in the 2×2 case. Some properties of the normal approximant and algorithms for finding it are given in A. Ruhe, *Closest normal matrix finally found!* BIT, 27 (1987) 585-598. The result in Problem IX.8.20 was proved, in the special case of the operator norm, by J. Phillips, *Nearest normal approximation for certain normal operators*, Proc. Amer. Math. Soc., 67 (1977) 236-240. The general result was proved in R. Bhatia, R. Horn, and F. Kittaneh, *Normal approximants to binormal operators*, Linear Algebra Appl., 147(1991) 169-179. An excellent survey of matrix approximation problems, with many references and applications, can be found in N.J. Higham, *Matrix nearness problems and applications*, in the collection *Applications of Matrix Theory*, Oxford University Press, 1989. A particularly striking application of Theorem IX.7.2 has been found in quantum chemistry. Given n linearly independent unit vectors $e_1, \ldots, e_n$ in an n-dimensional Hilbert space, what is the orthonormal basis $f_1, \ldots, f_n$ that is closest to the e_j, in the sense that $\Sigma \|e_j - f_j\|^2$ is minimal? The Gram-Schmidt procedure does not lead to such an orthonormal basis. The chemist P.O. Löwdin, *On the*

non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals, J. Chem. Phys., 18(1950) 365-374, found a procedure to obtain such a basis. The problem is clearly equivalent to that of finding a unitary matrix closest to an invertible matrix, in the Hilbert-Schmidt norm. Theorem IX.7.2 solves the problem for all unitarily invariant norms. The importance of such results is explained in J.A. Goldstein and M. Levy, *Linear algebra and quantum chemistry*, American Math. Monthly, 78 (1991) 710-718.

X

Perturbation of Matrix Functions

In earlier chapters we derived several inequalities that describe the variation of eigenvalues, eigenvectors, determinants, permanents, and tensor powers of a matrix. Similar problems for some other matrix functions are studied in this chapter.

X.1 Operator Monotone Functions

If a, b are positive real numbers, then it is easy to see that $|a^r - b^r| \geq |a-b|^r$ if $r \geq 1$, and $|a^r - b^r| \leq |a-b|^r$ if $0 \leq r \leq 1$. The inequalities in this section are extensions of these elementary inequalities to positive operators A, B. Instead of the power functions $f(t) = t^r$, $0 \leq r \leq 1$, we shall consider the more general class of operator monotone functions.

Theorem X.1.1 *Let f be an operator monotone function on $[0, \infty)$ such that $f(0) = 0$. Then for all positive operators A, B,*

$$\|f(A) - f(B)\| \leq f(\|A - B\|). \tag{X.1}$$

Proof. Since f is concave (Theorem V.2.5) and $f(0) = 0$, we have $f(a+b) \leq f(a) + f(b)$ for all nonnegative numbers a, b.

Let $\alpha = \|A - B\|$. Then $A - B \leq \alpha I$. Hence, $A \leq B + \alpha I$ and $f(A) \leq f(B + \alpha I)$. By the subadditivity property of f mentioned above, therefore, $f(A) \leq f(B) + f(\alpha)I$. Thus $f(A) - f(B) \leq f(\alpha)I$ and, by symmetry, $f(B) - f(A) \leq f(\alpha)I$. This implies that $|f(A) - f(B)| \leq f(\alpha)I$. Hence, $\|f(A) - f(B)\| \leq f(\alpha) = f(\|A - B\|)$. ■

Note the special consequence of the above theorem:

$$\|A^r - B^r\| \leq \|A - B\|^r, \quad 0 \leq r \leq 1 \tag{X.2}$$

for any two positive operators A, B. Note also that the argument in the above proof shows that

$$|||f(A) - f(B)||| \leq f(\|A - B\|)|||I|||$$

for every unitarily invariant norm.

Exercise X.1.2 *Show that for* 2×2 *positive matrices, the inequality* $\|A^{1/2} - B^{1/2}\|_2 \leq \|A - B\|_2^{1/2}$ *is not always valid. (It is false even when* $B = 0$.)

The inequality (X.2) can be rewritten in another form:

$$\|A^r - B^r\| \leq \|\, |A - B|^r \,\|, \quad 0 \leq r \leq 1. \tag{X.3}$$

This has a generalisation to all unitarily invariant norms. Once again, for this generalisation, it is convenient to consider the more general class of operator monotone functions.

Recall that every operator monotone function f on $[0, \infty)$ has an integral representation

$$f(t) = \gamma + \beta t + \int_0^\infty \frac{\lambda t}{\lambda + t} dw(\lambda), \tag{X.4}$$

where $\gamma = f(0)$, $\beta \geq 0$ and w is a positive measure such that $\int_0^\infty \frac{\lambda}{1+\lambda} dw(\lambda) < \infty$. (See (V.53).)

Theorem X.1.3 *Let f be an operator monotone function on $[0, \infty)$ such that $f(0) = 0$. Then for all positive operators A, B and for all unitarily invariant norms*

$$|||f(A) - f(B)||| \leq |||f(|A - B|)|||. \tag{X.5}$$

In the proof of the theorem, we will use the following lemma.

Lemma X.1.4 *Let X, Y be positive operators. Then*

$$|||(X + I)^{-1} - (X + Y + I)^{-1}||| \leq |||I - (Y + I)^{-1}|||$$

for every unitarily invariant norm.

Proof. Since $(X+I)^{-1} \leq I$, by Lemma V.1.5 we have $Y^{1/2}(X+I)^{-1}Y^{1/2} \leq Y$. Hence,

$$I - [Y^{1/2}(X + I)^{-1}Y^{1/2} + I]^{-1} \leq I - (Y + I)^{-1}.$$

Therefore, by the Weyl Monotonicity Principle,

$$\lambda_j^{\downarrow}(I - [Y^{1/2}(X+I)^{-1}Y^{1/2} + I]^{-1}) \leq \lambda_j^{\downarrow}(I - [Y+I]^{-1})$$

for all j. Note that $Y^{1/2}(X+I)^{-1}Y^{1/2}$ has the same eigenvalues as those of $(X+I)^{-\frac{1}{2}}Y(X+I)^{-\frac{1}{2}}$. So, the above inequality can also be written as

$$\lambda_j^{\downarrow}(I - [(X+I)^{-\frac{1}{2}}Y(X+I)^{-\frac{1}{2}} + I]^{-1}) \leq \lambda_j^{\downarrow}(I - [Y+I]^{-1}).$$

From the identity

$$\begin{aligned}(X+I)^{-1} &- (X+Y+I)^{-1} \\ &= (X+I)^{-\frac{1}{2}}\{I - [(X+I)^{-\frac{1}{2}}Y(X+I)^{-\frac{1}{2}} + I]^{-1}\}(X+I)^{-\frac{1}{2}}\end{aligned}$$

and the fact that $\|(X+I)^{-\frac{1}{2}}\| \leq 1$, we see that

$$\begin{aligned}&\lambda_j^{\downarrow}([X+I]^{-1} - [X+Y+I]^{-1}) \\ &\leq \lambda_j^{\downarrow}(I - [(X+I)^{-\frac{1}{2}}Y(X+I)^{-\frac{1}{2}} + I]^{-1}).\end{aligned}$$

Thus,

$$\lambda_j^{\downarrow}([X+I]^{-1} - [X+Y+I]^{-1}) \leq \lambda_j^{\downarrow}(I - [Y+I]^{-1})$$

for all j. This is more than what we need to prove the lemma. ∎

Proof of Theorem X.1.3: By the Fan Dominance Property (Theorem IV.2.2) it is sufficient to prove the inequality (X.5) for the special class of norms $\|\cdot\|_{(k)}$, $k = 1, 2, \ldots, n$.

We will first consider the case when $A - B$ is positive. Let $C = A - B$. We want to prove that

$$\|f(B+C) - f(B)\|_{(k)} \leq \|f(C)\|_{(k)}. \tag{X.6}$$

Let $\sigma_j = s_j(C)$, $j = 1, 2, \ldots, n$. Since σ_j are the eigenvalues of the positive operator C, we have

$$s_j(h(C)) = h(\sigma_j), \quad j = 1, \ldots, n,$$

for every monotonically increasing nonnegative function $h(t)$ on $[0, \infty)$. Thus

$$\|h(C)\|_{(k)} = \sum_{j=1}^{k} h(\sigma_j), \quad k = 1, \ldots, n,$$

for all such functions h.

Now, f has the representation (X.4) with $\gamma = 0$. The functions βt and $\frac{\lambda t}{\lambda + t}$ are nonnegative, monotonically increasing functions of t. Hence,

$$\begin{aligned} \|f(C)\|_{(k)} &= \sum_{j=1}^{k} f(\sigma_j) \\ &= \beta \sum_{j=1}^{k} \sigma_j + \int_0^\infty \sum_{j=1}^{k} \frac{\lambda \sigma_j}{\lambda + \sigma_j} dw(\lambda) \\ &= \beta \|C\|_{(k)} + \int_0^\infty \lambda \|C(C+\lambda I)^{-1}\|_{(k)} dw(\lambda). \end{aligned}$$

In the same way, we can obtain from the integral representation of f

$$\begin{aligned} &\|f(B+C) - f(B)\|_{(k)} \\ &\leq \beta \|C\|_{(k)} + \int_0^\infty \lambda \|(B+C)(B+C+\lambda I)^{-1} - B(B+\lambda I)^{-1}\|_{(k)} dw(\lambda). \end{aligned}$$

Thus, our assertion (X.6) will be proved if we show that for each $\lambda > 0$

$$\|(B+C)(B+C+\lambda I)^{-1} - B(B+\lambda I)^{-1}\|_{(k)} \leq \|C(C+\lambda I)^{-1}\|_{(k)}.$$

Now note that we can write

$$X(X+\lambda I)^{-1} = I - \left(\frac{X}{\lambda} + I\right)^{-1}.$$

So, the above inequality follows from Lemma X.1.4. This proves the theorem in the special case when $A - B$ is positive.

To prove the general case we will use the special case proved above and two simple facts. First, if X, Y are Hermitian with positive parts X^+, Y^+ in their respective Jordan decompositions, then the inequality $X \leq Y$ implies that $\|X^+\|_{(k)} \leq \|Y^+\|_{(k)}$ for all k. This is an immediate consequence of Weyl's Monotonicity Principle. Second, if X_1, X_2, Y_1, Y_2 are positive operators such that $X_1 X_2 = 0$, $Y_1 Y_2 = 0$, $\|X_1\|_{(k)} \leq \|Y_1\|_{(k)}$, and $\|X_2\|_{(k)} \leq \|Y_2\|_{(k)}$ for all k, then we have $\|X_1 + X_2\|_{(k)} \leq \|Y_1 + Y_2\|_{(k)}$ for all k. This can be easily seen using the fact that since X_1 and X_2 commute they can be simultaneously diagonalised, and so can Y_1, Y_2.

Now let A, B be any two positive operators. Since $A - B \leq (A-B)^+$, we have $A \leq B + (A-B)^+$, and hence $f(A) \leq f(B + (A-B)^+)$. From this we have

$$f(A) - f(B) \leq f(B + (A-B)^+) - f(B),$$

and, therefore, by the first observation above,

$$\|[f(A) - f(B)]^+\|_{(k)} \leq \|f(B + (A-B)^+) - f(B)\|_{(k)}$$

for all k. Then, using the special case of the theorem proved above we can conclude that

$$\|[f(A)-f(B)]^{+}\|_{(k)} \leq \|f([A-B]^{+})\|_{(k)}$$

for all k. Interchanging A and B, we have

$$\|[f(B)-f(A)]^{+}\|_{(k)} \leq \|f([B-A]^{+})\|_{(k)}$$

for all k. Now note that

$$\begin{aligned} f([A-B]^{+})f([B-A]^{+}) &= 0, \\ f([A-B]^{+})+f([B-A]^{+}) &= f(|A-B|), \\ [f(A)-f(B)]^{+}\,[f(B)-f(A)]^{+} &= 0, \\ [f(A)-f(B)]^{+}+[f(B)-f(A)]^{+} &= |f(A)-f(B)|. \end{aligned}$$

Thus, the two inequalities above imply that

$$\|f(A)-f(B)\|_{(k)} \leq \|f(|A-B|)\|_{(k)}$$

for all k. This proves the theorem. ∎

Exercise X.1.5 *Show that the conclusion of Theorem X.1.3 is valid for all nonnegative operator monotone functions on* $[0,\infty)$*; i.e., we can replace the condition* $f(0)=0$ *by the condition* $f(0)\geq 0$.

One should note two special corollaries of the theorem: we have for all positive operators A, B and for all unitarily invariant norms

$$|||A^r - B^r||| \leq |||\, |A-B|^r \,|||, \quad 0 \leq r \leq 1, \tag{X.7}$$

$$|||\log(I+A)-\log(I+B)||| \leq |||\log(I+|A-B|)|||. \tag{X.8}$$

Theorem X.1.6 *Let g be a continuous strictly increasing map of* $[0,\infty)$ *onto itself. Suppose that the inverse map g^{-1} is operator monotone. Then for all positive operators A, B and for all unitarily invariant norms, we have*

$$|||g(A)-g(B)||| \geq |||g(|A-B|)|||. \tag{X.9}$$

Proof. Let $f=g^{-1}$. Since f is operator monotone, it is concave by Theorem V.2.5. Hence g is convex. From Theorem X.1.3, with $g(A)$ and $g(B)$ in place of A and B, we have

$$|||A-B||| \leq |||f(|g(A)-g(B)|)|||.$$

This is equivalent to the weak majorisation

$$\{s_j(A-B)\} \prec_w \{s_j(f(|g(A)-g(B)|))\}.$$

Since f is monotone,

$$s_j(f(|g(A)-g(B)|)) = f(s_j(g(A)-g(B)))$$

for each j. So, we have

$$\{s_j(A-B)\} \prec_w \{f(s_j(g(A)-g(B)))\}.$$

Since g is convex and monotone, by Corollary II.3.4, we have from this

$$\{g(s_j(A-B))\} \prec_w \{s_j(g(A)-g(B))\}.$$

Since g is monotone, this is the same as saying that

$$\{s_j(g|A-B|)\} \prec_w \{s_j(g(A)-g(B))\},$$

and this, in turn, implies the inequality (X.9). ■

Two special corollaries that complement the inequalities (X.7) and (X.8) are worthy of note. For all positive operators A, B and for all unitarily invariant norms,

$$|||A^r - B^r||| \geq |||\,|A-B|^r|||, \text{ if } r \geq 1, \tag{X.10}$$

$$|||\exp A - \exp B||| \geq |||\exp(|A-B|) - I|||. \tag{X.11}$$

Exercise X.1.7 *Derive the inequality (X.10) from (X.7) using Exercise IV.2.8.*

Is there an inequality like (X.10) for Hermitian operators A, B? First note that if A is Hermitian, only positive integral powers of A are meaningfully defined. So, the question is whether $|||(A-B)^m|||$ can be bounded above by $|||A^m - B^m|||$. No such bound is possible if m is an even integer; for the choice $B = -A$, we have $A^m - B^m = 0$. For odd integers m, we do have a satisfactory answer.

Theorem X.1.8 *Let A, B be Hermitian operators. Then for all unitarily invariant norms, and for $m = 1, 2, \ldots,$*

$$|||(A-B)^{2m+1}||| \leq 2^{2m}|||A^{2m+1} - B^{2m+1}|||. \tag{X.12}$$

For the proof we need the following lemma.

Lemma X.1.9 *Let A, B be Hermitian operators, let X be any operator, and let m be any positive integer. Then, for $j = 1, 2, \ldots, m,$*

$$\begin{aligned} &|||A^{m+j}XB^{m-j+1} - A^{m-j+1}XB^{m+j}||| \\ \leq\; &|||A^{m+j+1}XB^{m-j} - A^{m-j}XB^{m+j+1}|||. \end{aligned} \tag{X.13}$$

Proof. By the arithmetic-geometric mean inequality proved in Theorem IX.4.5, we have

$$|||AXB||| \leq 1/2 \; |||A^2X + XB^2|||$$

for all operators X and Hermitian A, B. We will use this to prove the inequality (X.13). First consider the case $j = 1$. We have

$$\begin{aligned} & |||A^{m+1}XB^m - A^mXB^{m+1}||| \\ = \;& |||A(A^mXB^{m-1} - A^{m-1}XB^m)B||| \\ \leq \;& 1/2 \; |||A^2(A^mXB^{m-1} - A^{m-1}XB^m) \\ & \qquad + (A^mXB^{m-1} - A^{m-1}XB^m)B^2||| \\ \leq \;& 1/2 \; |||A^{m+2}XB^{m-1} - A^{m-1}XB^{m+2}||| \\ & \qquad + 1/2 \; |||A^{m+1}XB^m - A^mXB^{m+1}|||. \end{aligned}$$

Hence

$$|||A^{m+1}XB^m - A^mXB^{m+1}||| \leq |||A^{m+2}XB^{m-1} - A^{m-1}XB^{m+2}|||.$$

This shows that the inequality (X.13) is true for $j = 1$. The general case is proved by induction. Suppose that the inequality (X.13) has been proved for $j-1$ in place of j. Then using the arithmetic-geometric mean inequality, the triangle inequality, and this induction hypothesis, we have

$$\begin{aligned} & |||A^{m+j}XB^{m-j+1} - A^{m-j+1}XB^{m+j}||| \\ = \;& |||A(A^{m+j-1}XB^{m-j} - A^{m-j}XB^{m+j-1})B||| \\ \leq \;& 1/2 \; |||A^2(A^{m+j-1}XB^{m-j} - A^{m-j}XB^{m+j-1}) \\ & \qquad + (A^{m+j-1}XB^{m-j} - A^{m-j}XB^{m+j-1})B^2||| \\ \leq \;& 1/2 \; |||A^{m+j+1}XB^{m-j} - A^{m-j}XB^{m+j+1}||| \\ & \qquad + 1/2 \; |||A^{m+(j-1)}XB^{m-(j-1)+1} - A^{m-(j-1)+1}XB^{m+(j-1)}||| \\ \leq \;& 1/2 \; |||A^{m+j+1}XB^{m-j} - A^{m-j}XB^{m+j+1}||| \\ & \qquad + 1/2 \; |||A^{m+j}XB^{m-(j-1)} - A^{m-(j-1)}XB^{m+j}|||. \end{aligned}$$

This proves the desired inequality. ∎

Proof of Theorem X.1.8: Using the triangle inequality and a very special case of Lemma X.1.9, we have

$$\begin{aligned} & |||A^{2m}(A - B) + (A - B)B^{2m}||| \\ \leq \;& |||A^{2m+1} - B^{2m+1}||| + |||A^{2m}B - AB^{2m}||| \\ \leq \;& 2|||A^{2m+1} - B^{2m+1}|||. \end{aligned} \tag{X.14}$$

Let $C = A - B$ and choose an orthonormal basis x_j such that $Cx_j = \lambda_j x_j$, where $|\lambda_j| = s_j(C)$ for $j = 1, 2, \ldots, n$. Then, by the extremal representation

of the Ky Fan norms given in Problem III.6.6,

$$\begin{aligned}\|A^{2m}C + CB^{2m}\|_{(k)} &\geq \sum_{j=1}^{k} |\langle x_j, (A^{2m}C + CB^{2m})x_j\rangle| \\ &= \sum_{j=1}^{k} |\lambda_j| \{\langle x_j, A^{2m}x_j\rangle + \langle x_j, B^{2m}x_j\rangle\}.\end{aligned}$$

Now use the observation in Problem IX.8.14 and the convexity of the function $f(t) = t^{2m}$ to see that

$$\begin{aligned}\langle x_j,\ A^{2m}x_j\rangle + \langle x_j, B^{2m}x_j\rangle &\geq \|Ax_j\|^{2m} + \|Bx_j\|^{2m} \\ &\geq 2^{1-2m}(\|Ax_j\| + \|Bx_j\|)^{2m} \\ &\geq 2^{1-2m}\|Ax_j - Bx_j\|^{2m} \\ &= 2^{1-2m}|\lambda_j|^{2m}.\end{aligned}$$

We thus have

$$\begin{aligned}\|A^{2m}C + CB^{2m}\|_{(k)} &\geq \sum_{j=1}^{k} 2^{1-2m}|\lambda_j|^{2m+1} \\ &= 2^{1-2m}\|(A - B)^{2m+1}\|_{(k)}.\end{aligned}$$

Since this is true for all k, we have

$$|||A^{2m}C + CB^{2m}||| \geq 2^{1-2m}|||(A - B)^{2m+1}|||$$

for all unitarily invariant norms. Combining this with (X.14) we obtain the inequality (X.12). ■

Observe that when $B = -A$, the two sides of (X.12) are equal.

X.2 The Absolute Value

In Section VII.5 we obtained bounds for $|||\ |A| - |B|\ |||$ in terms of $|||A - B|||$. More such bounds are obtained in this section. Since $|A| = (A^*A)^{1/2}$, results for the square root function obtained in the preceding section are useful here.

Theorem X.2.1 *Let A, B be any two operators. Then, for every unitarily invariant norm,*

$$|||\ |A| - |B|\ ||| \leq \sqrt{2}(|||A + B|||\ |||A - B|||)^{1/2}. \qquad \text{(X.15)}$$

Proof. From the inequality (X.7) we have

$$|||\ |A| - |B|\ ||| \leq |||\ |A^*A - B^*B|^{1/2}|||. \qquad \text{(X.16)}$$

Note that

$$A^*A - B^*B = 1/2\ \{(A+B)^*(A-B) + (A-B)^*(A+B)\}. \qquad \text{(X.17)}$$

Hence, by Theorem III.5.6, we can find unitaries U and V such that

$$|A^*A - B^*B| \leq 1/2\ \{U|(A+B)^*(A-B)|U^* + V|(A-B)^*(A+B)|V^*\}.$$

Since the square root function is operator monotone, this operator inequality is preserved if we take square roots on both sides. Since every unitarily invariant norm is monotone, this shows that

$$\begin{aligned} \||\, |A^*A - B^*B|^{1/2}\||^2 \quad &\leq \quad 1/2\ \||[U|(A+B)^*(A-B)|U^* \\ &\qquad + V|(A-B)^*(A+B)|V^*]^{1/2}\||^2. \end{aligned}$$

By the result of Problem IV.5.6, we have

$$\||\, |X+Y|^{1/2}\||^2 \leq 2(\||\, |X|^{1/2}\||^2 + \||\, |Y|^{1/2}\||^2)$$

for all X, Y. Hence,

$$\begin{aligned} \||\, |A^*A - B^*B|^{1/2}\||^2 \quad &\leq \quad \||\, |(A+B)^*(A-B)|^{1/2}\||^2 \\ &\qquad + \||\, |(A-B)^*(A+B)|^{1/2}\||^2. \end{aligned}$$

By the Cauchy-Schwarz inequality (IV.44), the right-hand side is bounded above by $2\||A+B\||\ \||A-B\||$. Thus the inequality (X.15) now follows from (X.16). ∎

Example X.2.2 *Let*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then

$$\|\, |A| - |B|\, \|_1 = 2, \quad \|A+B\|_1 = \|A-B\|_1 = \sqrt{2}.$$

So, for the trace norm the inequality (X.15) is sharp.

An improvement of this inequality is possible for special norms.

Theorem X.2.3 *Let A, B be any two operators. Then for every Q-norm (and thus, in particular, for every Schatten p-norm with $p \geq 2$) we have*

$$\|\, |A| - |B|\, \|_Q \leq (\|A+B\|_Q \|A-B\|_Q)^{1/2}. \qquad \text{(X.18)}$$

Proof. By the definition of Q-norms, the inequality (X.18) is equivalent to the assertion that

$$\||(|A| - |B|)^2\|| \leq \||\, |A+B|^2\||^{1/2} \||\, |A-B|^2\||^{1/2} \qquad \text{(X.19)}$$

for all unitarily invariant norms. From the inequality (X.10) we have

$$|||(|A|-|B|)^2||| \leq |||\, |A|^2 - |B|^2||| = |||A^*A - B^*B|||.$$

Using the identity (X.17), we see that

$$|||A^*A - B^*B||| \leq 1/2\ \{|||(A+B)^*(A-B)||| + |||(A-B)^*(A+B)|||\}.$$

Now, using the Cauchy-Schwarz inequality given in Problem IV.5.7, we see that each of the two terms in the brackets above is dominated by

$$|||\, |A+B|^2|||^{1/2}|||\, |A-B|^2|||^{1/2}.$$

This proves the inequality (X.19). ■

Theorem X.2.4 *Let A, B be any two operators. Then, for all Schatten p-norms with $1 \leq p \leq 2$,*

$$\|\, |A| - |B|\, \|_p \leq 2^{\frac{1}{p}-\frac{1}{2}}(\|A+B\|_p\|A-B\|_p)^{1/2}. \tag{X.20}$$

Proof. Let

$$\|X\|_p := \left(\Sigma\ s_j^p(X)\right)^{1/p}, \quad \text{for all } p > 0.$$

When $p \geq 1$, these are the Schatten p-norms. When $0 < p < 1$, this defines a quasi-norm. Instead of the triangle inequality, we have

$$\|X+Y\|_p \leq 2^{1/p-1}(\|X\|_p + \|Y\|_p), \quad 0 < p < 1. \tag{X.21}$$

(See Problems IV.5.1 and IV.5.6.) Note that for all positive real numbers r and p, we have

$$\|\, |X|^r \|_p = \|X\|_{rp}^r. \tag{X.22}$$

Thus, the inequality (X.7), restricted to the p-norms, gives for all positive operators A, B

$$\|A^r - B^r\|_p \leq \|A-B\|_{rp}^r \quad \text{for } \ 0 < r \leq 1,\ 1 \leq p \leq \infty. \tag{X.23}$$

Hence, for any two operators A, B,

$$\|\, |A| - |B|\, \|_p \leq \|A^*A - B^*B\|_{p/2}^{1/2}, \quad 1 \leq p \leq \infty.$$

Now use the identity (X.17) and the property (X.21) to see that, for $1 \leq p \leq 2$,

$$\|A^*A - B^*B\|_{p/2} \leq 2^{2/p-2}\{\|(A+B)^*(A-B)\|_{p/2} + \|(A-B)^*(A+B)\|_{p/2}\}.$$

From the relation (X.22) and the Cauchy-Schwarz inequality (IV.44), it follows that each of the two terms in the brackets is dominated by $\|A+B\|_p\|A-B\|_p$. Hence

$$\|A^*A - B^*B\|_{p/2} \leq 2^{2/p-1}\|A+B\|_p\|A-B\|_p$$

for $1 \leq p \leq 2$. This proves the theorem. ■

The example given in X.2.2 shows that, for each $1 \leq p \leq 2$, the inequality (X.20) is sharp.

In Section VII.5 we saw that

$$\| \, |A| - |B| \, \|_2 \leq \sqrt{2} \, \|A - B\|_2 \tag{X.24}$$

for any two operators A, B. Further, if both A and B are normal, the factor $\sqrt{2}$ can be replaced by 1. Can one prove a similar inequality for the operator norm instead of the Hilbert-Schmidt norm? Of course, we have from (X.24)

$$\| \, |A| - |B| \, \| \leq \sqrt{2n} \, \|A - B\| \tag{X.25}$$

for all operators A, B on an n-dimensional Hilbert space $\mathcal{H}$. It is known that the factor $\sqrt{2n}$ in the above inequality can be replaced by a factor $c_n = O(\log \, n)$; and even when both A, B are Hermitian, such a factor is necessary. (See the Notes at the end of the chapter.)

In a slightly different vein, we have the following theorem.

Theorem X.2.5 *(T. Kato) For any two operators A, B we have*

$$\| \, |A| - |B| \, \| \leq \frac{2}{\pi} \|A - B\| \left(2 + \log \, \frac{\|A\| + \|B\|}{\|A - B\|} \right). \tag{X.26}$$

Proof. The square root function has an integral representation (V.4); this says that

$$t^{1/2} = \frac{1}{\pi} \int_0^\infty \frac{t}{\lambda + t} \lambda^{-1/2} d\lambda.$$

We can rewrite this as

$$t^{1/2} = \frac{1}{\pi} \int_0^\infty [\lambda^{-1/2} - \lambda^{1/2}(\lambda + t)^{-1}] d\lambda.$$

Using this, we have

$$|A| - |B| = \frac{1}{\pi} \int_0^\infty \lambda^{1/2} [(|B|^2 + \lambda)^{-1} - (|A|^2 + \lambda)^{-1}] d\lambda. \tag{X.27}$$

We will estimate the norm of this integral by splitting it into three parts. Let

$$\alpha = \|A - B\|^2, \;\; \beta = (\|A\| + \|B\|)^2.$$

Now note that if X, Y are two positive operators, then $-Y \le X - Y \le X$, and hence $\|X - Y\| \le \max(\|X\|, \|Y\|)$. Using this we see that

$$\|\int_0^\alpha \lambda^{1/2}[(|B|^2+\lambda)^{-1} - (|A|^2+\lambda)^{-1}]d\lambda\| \le \int_0^\alpha \lambda^{-1/2}d\lambda = 2\alpha^{1/2} = 2\|A-B\|. \tag{X.28}$$

From the identity

$$(|B|^2+\lambda)^{-1} - (|A|^2+\lambda)^{-1} = (|B|^2+\lambda)^{-1}(|A|^2-|B|^2)(|A|^2+\lambda)^{-1} \tag{X.29}$$

and the identity (X.17), we see that

$$\begin{aligned} \|(|B|^2+\lambda)^{-1} - (|A|^2+\lambda)^{-1}\| &\le \lambda^{-2}\|A+B\|\,\|A-B\| \\ &\le \lambda^{-2}\beta^{1/2}\|A-B\|. \end{aligned}$$

Hence,

$$\|\int_\beta^\infty \lambda^{1/2}[(|B|^2+\lambda)^{-1} - (|A|^2+\lambda)^{-1}]d\lambda\| \le 2\|A-B\|. \tag{X.30}$$

Since $A^*A - B^*B = B^*(A-B) + (A^*-B^*)A$, from (X.29) we have

$$\begin{aligned} &(|B|^2+\lambda)^{-1} - (|A|^2+\lambda)^{-1} \\ &= (|B|^2+\lambda)^{-1}B^*(A-B)(|A|^2+\lambda)^{-1} \\ &+ (|B|^2+\lambda)^{-1}(A^*-B^*)A(|A|^2+\lambda)^{-1}. \end{aligned} \tag{X.31}$$

Note that

$$\begin{aligned} \|(|B|^2+\lambda)^{-1}B^*\| &= \|B(|B|^2+\lambda)^{-1}\| \\ &= \|\,|B|(|B|^2+\lambda)^{-1}\| \le \frac{1}{2\lambda^{1/2}}, \end{aligned}$$

since the maximum value of the function $f(t) = \frac{t}{t^2+\lambda}$ is $\frac{1}{2\lambda^{1/2}}$. By the same argument,

$$\|A(|A|^2+\lambda)^{-1}\| \le \frac{1}{2\lambda^{1/2}}.$$

So, from (X.31) we obtain

$$\|(|B|^2+\lambda)^{-1} - (|A|^2+\lambda)^{-1}\| \le \lambda^{-3/2}\|A-B\|.$$

Hence

$$\|\int_\alpha^\beta \lambda^{1/2}[(|B|^2+\lambda)^{-1}-(|A|^2+\lambda)^{-1}]d\lambda\| \leq \|A-B\|\int_\alpha^\beta \lambda^{-1}d\lambda$$

$$= \|A-B\| \log\frac{\beta}{\alpha} = 2\|A-B\|\log\frac{\|A\|+\|B\|}{\|A-B\|}. \qquad \text{(X.32)}$$

Combining (X.27), (X.28), (X.30), and (X.32), we obtain the inequality (X.26). ∎

X.3 Local Perturbation Bounds

Inequalities obtained above are global, in the sense that they are valid for all pairs of operators A, B. Some special results can be obtained if B is restricted to be close to A, or when both are restricted to be away from 0. It is possible to derive many interesting inequalities by using only elementary calculus on normed spaces. A quick review of the basic concepts of the Fréchet differential calculus that are used below is given in the Appendix.

Let f be any continuously differentiable map on an open interval I. Then the map that f induces on the set of Hermitian matrices whose spectrum is contained in I is Fréchet differentiable. This has been proved in Theorem V.3.3, and an explicit formula for the derivative is also given there. For each A, the derivative $Df(A)$ is a linear operator on the space of all Hermitian matrices. The norm of this operator is defined as

$$\|Df(A)\| = \sup_{\|B\|=1}\|Df(A)(B)\|. \qquad \text{(X.33)}$$

More generally, any unitarily invariant norm on Hermitian matrices leads to a corresponding norm for the linear operator $Df(A)$; we denote this as

$$|||Df(A)||| = \sup_{|||B|||=1}|||Df(A)(B)|||. \qquad \text{(X.34)}$$

For some special functions f, we will find upper bounds for these quantities. Among the functions we consider are operator monotone functions on $(0,\infty)$. The square root function $f(t)=t^{1/2}$ is easier to handle, and since it is especially important, it is worthwhile to deal with it separately.

Theorem X.3.1 *Let $f(t) = t^{1/2}$, $0 < t < \infty$. Then for every positive operator A, and for every unitarily invariant norm,*

$$|||Df(A)||| \leq 1/2\ \|A^{-1}\|^{1/2}. \qquad \text{(X.35)}$$

Proof. The function $g(t)=t^2$, $0<t<\infty$ is the inverse of f. So, by the chain rule of differentiation, $Df(A) = [Dg(f(A))]^{-1}$ for every positive

operator A. Note that $Dg(A)(X) = AX + XA$, for every X. So

$$[Dg(f(A))](X) = A^{1/2}X + XA^{1/2}.$$

If A has eigenvalues $\alpha_1 \geq \cdots \geq \alpha_n > 0$, then $\operatorname{dist}(\sigma(A^{1/2}), \sigma(-A^{1/2})) = 2\alpha_n^{1/2} = 2\|A^{-1}\|^{-1/2}$. Hence, by Theorem VII.2.12,

$$|||[Dg(f(A))]^{-1}||| \leq 1/2 \, \|A^{-1}\|^{1/2}.$$

This proves the theorem. ∎

Exercise X.3.2 *Let $f \in C^1(I)$ and let f' be the derivative of f. Show that*

$$\|f'(A)\| = \|Df(A)(I)\| \leq \|Df(A)\|. \tag{X.36}$$

Thus, for the function $f(t) = t^{1/2}$ on $(0, \infty)$,

$$\|Df(A)\| = \|f'(A)\| \tag{X.37}$$

for all positive operators A.

Theorem X.3.3 *Let φ be the map that takes an invertible operator A to its absolute value $|A|$. Then, for every unitarily invariant norm,*

$$|||D\varphi(A)||| \leq \operatorname{cond}(A) = \|A^{-1}\| \, \|A\|. \tag{X.38}$$

Proof. Let $g(A) = A^*A$. Then $Dg(A)(B) = A^*B + B^*A$. Hence $|||Dg(A)||| \leq 2\|A\|$. The map φ is the composite fg, where $f(A) = A^{1/2}$. So, by the chain rule, $D\varphi(A) = Df(g(A))Dg(A) = Df(A^*A)Dg(A)$. Hence

$$|||D\varphi(A)||| \leq |||Df(A^*A)||| \; |||Dg(A)|||.$$

The first term on the right is bounded by $\frac{1}{2}\|A^{-1}\|$ by Theorem X.3.1, and the second by $2\|A\|$. This proves the theorem. ∎

The following theorem generalises Theorem X.3.1.

Theorem X.3.4 *Let f be an operator monotone function on $(0, \infty)$. Then, for every unitarily invariant norm,*

$$|||Df(A)||| \leq \|f'(A)\| \tag{X.39}$$

for all positive operators A.

Proof. Use the integral representation (V.49) to write

$$f(t) = \alpha + \beta t + \int_0^\infty \left(\frac{\lambda}{\lambda^2 + 1} - \frac{1}{\lambda + t} \right) d\mu(\lambda),$$

where α, β are real numbers, $\beta \geq 0$, and μ is a positive measure. Thus

$$f(A) = \alpha I + \beta A + \int_0^\infty \left[\frac{\lambda}{\lambda^2+1} I - (\lambda + A)^{-1} \right] d\mu(\lambda).$$

Using the fact that, for the function $g(A) = A^{-1}$ we have $Dg(A)(B) = -A^{-1}BA^{-1}$, we obtain from the above expression

$$Df(A)(B) = \beta B + \int_0^\infty (\lambda + A)^{-1} B (\lambda + A)^{-1} d\mu(\lambda).$$

Hence

$$|||Df(A)||| \leq \beta + \int_0^\infty \|(\lambda + A)^{-1}\|^2 d\mu(\lambda). \tag{X.40}$$

From the integral representation we also have

$$f'(t) = \beta + \int_0^\infty \frac{1}{(\lambda + t)^2} d\mu(\lambda).$$

Hence

$$\|f'(A)\| = \|\beta I + \int_0^\infty (\lambda + A)^{-2} d\mu(\lambda)\|. \tag{X.41}$$

If A has eigenvalues $\alpha_1 \geq \cdots \geq \alpha_n$, then since $\beta \geq 0$, the right-hand sides of both (X.40) and (X.41) are equal to

$$\beta + \int_0^\infty (\lambda + \alpha_n)^{-2} d\mu(\lambda).$$

This proves the theorem. ■

Exercise X.3.5 *Let f be an operator monotone function on* $(0, \infty)$. *Show that*

$$\|Df(A)\| = \|Df(A)(I)\| = \|f'(A)\|.$$

Once we have estimates for the derivative $Df(A)$, we can obtain bounds for $|||f(A) - f(B)|||$ when B is close to A. These bounds are obtained using Taylor's Theorem and the mean value theorem.

Using Taylor's Theorem, we obtain from Theorems X.3.3 and X.3.4 above the following.

Theorem X.3.6 *Let A be an invertible operator. Then for every unitarily invariant norm*

$$|||\ |A| - |B|\ ||| \leq \text{cond}(A)|||A - B||| + O(|||A - B|||^2) \tag{X.42}$$

for all B close to A.

Theorem X.3.7 *Let f be an operator monotone function on $(0, \infty)$. Let A be any positive operator. Then for every unitarily invariant norm*

$$|||f(A) - f(B)||| \leq \|f'(A)\|\ |||A - B||| + O(|||A - B|||^2) \tag{X.43}$$

for all positive operators B close to A.

For the functions $f(t) = t^r$, $0 < r < 1$, we have from this

$$|||A^r - B^r||| \leq r\|A^{-1}\|^{1-r}|||A - B||| + O(|||A - B|||^2). \tag{X.44}$$

The use of the mean value theorem is illustrated in the proof of the following theorem.

Theorem X.3.8 *Let f be an operator monotone function on $(0, \infty)$ and let A, B be two positive operators that are bounded below by a; i.e., $A \geq aI$ and $B \geq aI$ for the positive number a. Then for every unitarily invariant norm*

$$|||f(A) - f(B)||| \leq f'(a)|||A - B|||. \tag{X.45}$$

Proof. Use the integral representation of f as in the proof of Theorem X.3.4. We have

$$f'(A) = \beta I + \int_0^\infty (\lambda + A)^{-2} d\mu(\lambda).$$

If $A \geq aI$, then

$$f'(A) \leq \beta I + [\int_0^\infty (\lambda + a)^{-2} d\mu(\lambda)]I = f'(a)I.$$

Let $A(t) = (1 - t)A + tB$, $0 \leq t \leq 1$. If A and B are bounded below by a, then so is $A(t)$. Hence, using the mean value theorem, the inequality (X.39), and the above observation, we have

$$\begin{aligned} |||f(A) - f(B)||| &\leq \sup_{0 \leq t \leq 1} |||Df(A(t))(A'(t))||| \\ &\leq \sup_{0 \leq t \leq 1} \|f'(A(t))\|\ |||A'(t)||| \\ &\leq f'(a)|||A - B|||. \end{aligned}$$

■

A special corollary is the inequality

$$|||A^r - B^r||| \leq r a^{r-1} |||A - B|||, \;\; 0 < r < 1, \tag{X.46}$$

valid for operators A, B such that $A \geq aI$ and $B \geq aI$ for some positive number a.

Other inequalities of this type are discussed in the Problems section.

Let $A = UP$ be the polar decomposition of an invertible matrix A. Then $P = |A|$ and $U = AP^{-1}$. Using standard rules of differentiation, one can obtain from this, expressions for the derivative of the map $A \to U$, and then obtain perturbation bounds for this map in the same way as was done for the map $A \to |A|$. There is, however, a more effective and simpler way of doing this.

The advantage of this new method, explained below, is that it also works for other decompositions like the QR decomposition, where explicit formulae for the two factors are not known. For this added power there is a small cost to be paid. The slightly more sophisticated notion of differentiation on a manifold of matrices has to be used. We have already used similar ideas in Chapter 6.

In the space $\mathbf{M(n)}$ of $n \times n$ matrices, let $\mathbf{GL(n)}$ be the set of all invertible matrices, $\mathbf{U(n)}$ the set of all unitary matrices, and $\mathbf{P(n)}$ the set of all positive (definite) matrices. All three are differentiable manifolds. The set $\mathbf{GL(n)}$ is an open subset of $\mathbf{M(n)}$, and hence the tangent space to $\mathbf{GL(n)}$ at each of its points is the space $\mathbf{M(n)}$. The tangent space to $\mathbf{U(n)}$ at the point I, written as $T_I\mathbf{U(n)}$, is the space $\mathcal{K}(\mathbf{n})$ of all skew-Hermitian matrices. This has been explained in Section VI.4. The tangent space at any other point U of $\mathbf{U(n)}$ is $T_U\mathbf{U(n)} = U \cdot \mathcal{K}(\mathbf{n}) = \{US : S \in \mathcal{K}(\mathbf{n})\}$. Let $\mathcal{H}(\mathbf{n})$ be the space of all Hermitian matrices. Both $\mathcal{H}(\mathbf{n})$ and $\mathcal{K}(\mathbf{n})$ are real vector spaces and $\mathcal{H}(\mathbf{n}) = i\mathcal{K}(\mathbf{n})$. The set $\mathbf{P(n)}$ is an open subset of $\mathcal{H}(\mathbf{n})$, and hence, the tangent space to $\mathbf{P(n)}$ at each of its points is $\mathcal{H}(\mathbf{n})$.

The polar decomposition gives a differentiable map Φ from $\mathbf{GL(n)}$ onto $\mathbf{U(n)} \times \mathbf{P(n)}$. This is the map $\Phi(A) = (\Phi_1(A), \; \Phi_2(A)) = (U, P)$, where the invertible matrix A has the polar decomposition $A = UP$. Earlier in this section we called $\Phi_2(A)$ just $\varphi(A)$ and evaluated its Fréchet derivative. An explicit formula for the derivative $D\Phi_1(A)$ is obtained below. This map is a linear map from $\mathbf{M(n)}$, the tangent space to $\mathbf{GL(n)}$, into the space $U \cdot \mathcal{K}(\mathbf{n})$, the tangent space to $\mathbf{U(n)}$ at the point U.

The main idea of the proof below is simple. Let Ψ be the map from $\mathbf{U(n)} \times \mathbf{P(n)}$ to $\mathbf{GL(n)}$ that is the inverse to Φ; i.e., $\Psi(U, P) = UP$. This is a much simpler object to handle, since it is just a product map. We can calculate the derivative of this map and then use the inverse function theorem to get the derivative of the map Φ.

Theorem X.3.9 *Let Φ_1 be the map from $\mathbf{GL(n)}$ into $\mathbf{U(n)}$ that takes an invertible matrix to the unitary part in its polar decomposition, $\Phi_1(UP) = U$. Then for each $X \in \mathbf{M(n)}$, the value of the derivative $D\Phi_1(UP)$ at the*

point UX is given by the formula

$$[D\Phi_1(UP)](UX) = 2U \int_0^\infty e^{-tP}(i\,\mathrm{Im}\,X)e^{-tP}dt. \tag{X.47}$$

Proof. The domain of the linear map $D\Psi(U,P)$ is the tangent space to the manifold $\mathbf{U}(\mathbf{n})\times\mathbf{P}(\mathbf{n})$ at the point (U,P). This space is $(U\cdot\mathcal{K}(\mathbf{n}),\mathcal{H}(\mathbf{n}))$. The range of $D\Psi(U,P)$ is the tangent space to $\mathbf{GL}(\mathbf{n})$ at UP. This space is $\mathbf{M}(\mathbf{n})$. We will use the decomposition $\mathbf{M}(\mathbf{n}) = U\cdot\mathcal{K}(\mathbf{n}) + U\cdot\mathcal{H}(\mathbf{n})$ that arises from the Cartesian decomposition. By the definition of the derivative, we have

$$\begin{aligned}
[D\Psi(U,P)](US,H) &= \left.\frac{d}{dt}\right|_{t=0} \Psi(Ue^{tS}, P+tH) \\
&= \left.\frac{d}{dt}\right|_{t=0} Ue^{tS}(P+tH) \\
&= USP + UH
\end{aligned}$$

for all $S\in\mathcal{K}(\mathbf{n})$ and $H\in\mathcal{H}(\mathbf{n})$.

The derivative $D\Phi(UP)$ is a linear map from $\mathbf{M}(\mathbf{n})$ onto $(U\cdot\mathcal{K}(\mathbf{n}),\mathcal{H}(\mathbf{n}))$. Suppose

$$[D\Phi(UP)](UX) = (UM,N).$$

Since $\Phi = \Psi^{-1}$, from the two equations above we see that

$$UX = [D\Phi(UP)]^{-1}(UM,N) = [D\Psi(U,P)](UM,N) = UMP + UN.$$

Hence,

$$X = MP + N.$$

Our task now is to find M from this equation. Note that M is skew-Hermitian and N Hermitian. Hence, from the above equation, we obtain

$$MP + PM = X - X^* = 2i\,\mathrm{Im}\,X.$$

This equation was studied in Chapter 7. From Theorem VII.2.3 we have its solution

$$M = 2\int_0^\infty e^{-tP}(i\,\mathrm{Im}\,X)e^{-tP}\,dt.$$

This gives us the expression (X.47). ■

Corollary X.3.10 *For every unitarily invariant norm we have*

$$|||D\Phi_1(UP)||| = \|P^{-1}\|. \tag{X.48}$$

Proof. Using (X.47) and properties of unitarily invariant norms, we see that

$$|||D\Phi_1(UP)(UX)||| \le 2\int_0^\infty \|e^{-tP}\|\ |||X|||\ \|e^{-tP}\|dt.$$

If P has eigenvalues $\alpha_1 \ge \cdots \ge \alpha_n$, then $\|e^{-tP}\| = e^{-t\alpha_n}$. So,

$$\begin{aligned} |||D\Phi_1(UP)(UX)||| &\le 2\int_0^\infty e^{-2t\alpha_n}|||X|||dt \\ &= \alpha_n^{-1}|||X||| = \|P^{-1}\|\ |||X|||. \end{aligned}$$

Hence,

$$|||D\Phi_1(UP)||| = \sup_{|||X|||=1} |||D\Phi_1(UP)(X)||| \le \|P^{-1}\|.$$

The choice $X = ivv^*/|||vv^*|||$, where v is an eigenvector of P belonging to the eigenvalue α_n, shows that the last inequality is in fact an equality. ■

Two corollaries follow; the first one is obtained using the mean value theorem and the second one using Taylor's Theorem.

Corollary X.3.11 *Let A_0, A_1 be two elements of* $\mathbf{GL(n)}$, *and let U_0, U_1 be the unitary factors in their polar decompositions. Suppose that the line segment $A(t) = (1-t)A_0 + tA_1$, $0 \le t \le 1$, lies inside* $\mathbf{GL(n)}$. *Then, for every unitarily invariant norm*

$$|||U_0 - U_1||| \le \max_{0\le t\le 1} \|A(t)^{-1}\| \cdot |||A_0 - A_1|||. \tag{X.49}$$

Corollary X.3.12 *Let A_0 be an invertible matrix with polar decomposition $A_0 = U_0P_0$. Then, for a matrix $A = UP$ in a neighbourhood of A_0, we have*

$$|||U_0 - U||| \le \|A_0^{-1}\|\ |||A_0 - A||| + O(|||A_0 - A|||^2). \tag{X.50}$$

Exercise X.3.13 *From the proof of Theorem X.3.9 one can also extract a bound for the derivative of the map $A \to |A|$. What does this give? Compare it with the result of Theorem X.3.3.*

Let us see now how this method works for a perturbation analysis of the QR decomposition.

Let $\mathbf{\Delta_+(n)}$ be the set of all upper triangular matrices with positive diagonal entries. Each element A of $\mathbf{GL(n)}$ has a unique factoring $A = QR$, where $Q \in \mathbf{U(n)}$ and $R \in \mathbf{\Delta_+(n)}$. Thus the QR decomposition gives rise to an invertible map Φ from $\mathbf{GL(n)}$ onto $\mathbf{U(n)} \times \mathbf{\Delta_+(n)}$. Let $\mathbf{\Delta_{re}(n)}$ be the set of all upper triangular matrices with real diagonal entries. This is a real vector space, and $\mathbf{\Delta_+(n)}$ is an open set in it. Thus the tangent space to $\mathbf{\Delta_+(n)}$, at any of its points, is $\mathbf{\Delta_{re}(n)}$. For each $A = QR$ in $\mathbf{GL(n)}$

the derivative $D\Phi(A)$ is a linear map from the vector space $\mathbf{M(n)}$ onto the vector space $(Q \cdot \mathcal{K}(\mathbf{n}), \mathbf{\Delta_{re}(n)})$. We want to calculate the norm of this.

First note that the spaces $\mathcal{K}(\mathbf{n})$ and $\mathbf{\Delta_{re}(n)}$ are complementary to each other in $\mathbf{M(n)}$. We have a vector space decomposition

$$\mathbf{M(n)} = \mathcal{K}(\mathbf{n}) + \mathbf{\Delta_{re}(n)}. \tag{X.51}$$

Every matrix X splits as $X = K + T$ in this decomposition; the entries of X, K and T are related as follows:

$$\begin{array}{lcll} k_{jj} & = & i \operatorname{Im} x_{jj} & \text{for all } j, \\ k_{ij} & = & -\bar{x}_{ji} & \text{for } j > i, \\ k_{ij} & = & x_{ij} & \text{for } i > j, \\ t_{jj} & = & \operatorname{Re} x_{jj} & \text{for all } j, \\ t_{ij} & = & x_{ij} + \bar{x}_{ji} & \text{for } j > i, \\ t_{ij} & = & 0 & \text{for } i > j. \end{array} \tag{X.52}$$

Exercise X.3.14 *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the complementary projection operators in $\mathbf{M(n)}$ corresponding to the decomposition (X.51). Show that*

$$\|\mathcal{P}_1\|_2 = \|\mathcal{P}_2\|_2 = \sqrt{2},$$

where $\|\mathcal{P}_j\|_2 = \sup_{\|X\|_2=1} \|\mathcal{P}_j X\|_2$, and $\|\cdot\|_2$ stands for the Frobenius (Hilbert-Schmidt) norm.

Now let Ψ be the map from $\mathbf{U(n)} \times \mathbf{\Delta_+(n)}$ onto $\mathbf{M(n)}$ defined as $\Psi(Q,R) = QR$. Then Ψ and Φ are inverse to each other. The derivative $D\Psi(Q,R)$ is a linear map whose domain is the tangent space to the manifold $\mathbf{U(n)} \times \mathbf{\Delta_+(n)}$ at the point (Q,R). This space is $(Q \cdot \mathcal{K}(\mathbf{n}),\ \mathbf{\Delta_{re}(n)})$. Its range is the space $\mathbf{M(n)} = Q \cdot \mathcal{K}(\mathbf{n}) + Q \cdot \mathbf{\Delta_{re}(n)}$. By the definition of the derivative, we have

$$\begin{aligned} [D\Psi(Q,R)](QK,T) &= \left.\frac{d}{dt}\right|_{t=0} \Psi(Qe^{tK}, R+tT) \\ &= \left.\frac{d}{dt}\right|_{t=0} Qe^{tK}(R+tT) \\ &= QKR + QT \end{aligned}$$

for all $K \in \mathcal{K}(\mathbf{n})$ and $T \in \mathbf{\Delta_{re}(n)}$.

The derivative $D\Phi(QR)$ is a linear map from $\mathbf{M(n)}$ onto $Q \cdot \mathcal{K}(\mathbf{n}) + \mathbf{\Delta_{re}(n)}$. Suppose

$$[D\Phi(QR)](QX) = (QM, N),$$

where $M \in \mathcal{K}(\mathbf{n})$ and $N \in \mathbf{\Delta_{re}(n)}$. Then we must have

$$QX = [D\Phi(QR)]^{-1}(QM,N) = [D\Psi(Q,R)](QM,N) = QMR + QN.$$

Hence

$$X = MR + N.$$

So, we have the same kind of equation as we had in the analysis of the polar decomposition. There is one vital difference, however. There, the matrices M, N were skew-Hermitian and Hermitian, respectively, and instead of the upper triangular factor R we had the positive factor P. So, taking adjoints, we could eliminate N and get another equation that we could solve explicitly. We cannot do that here. But there is another way out. We have from the above equation

$$XR^{-1} = M + NR^{-1}. \tag{X.53}$$

Here $M \in \mathcal{K}(\mathbf{n})$; and both N and R^{-1} are in $\mathbf{\Delta_{re}(n)}$, and hence so is their product NR^{-1}. Thus the equation (X.53) is nothing but the decomposition of XR^{-1} with respect to the vector space decomposition (X.51). In this way, we now know M and N explicitly. We thus have the following theorem.

Theorem X.3.15 *Let Φ_1, Φ_2 be the maps from $\mathbf{GL(n)}$ into $\mathbf{U(n)}$ and $\mathbf{\Delta_+(n)}$ that take an invertible matrix to the unitary and the upper triangular factors in its QR decomposition. Then for each $X \in \mathbf{M(n)}$, the derivatives $D\Phi_1(QR)$ and $D\Phi_2(QR)$ evaluated at the point QX are given by the formulae*

$$[D\Phi_1(QR)](QX) = Q\mathcal{P}_1(XR^{-1}),$$

$$[D\Phi_2(QR)](QX) = \mathcal{P}_2(XR^{-1})R,$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the complementary projection operators in $\mathbf{M(n)}$ corresponding to the decomposition (X.51).

Using the result of Exercise X.3.14, we obtain the first corollary below. Then the next two corollaries are obtained using the mean value theorem and Taylor's Theorem.

Corollary X.3.16 *Let Φ_1, Φ_2 be the maps that take an invertible matrix A to the Q and R factors in its QR decomposition. Then*

$$\begin{aligned} \|D\Phi_1(A)\|_2 &\leq \sqrt{2}\,\|A^{-1}\|, \\ \|D\Phi_2(A)\|_2 &\leq \sqrt{2}\,\mathrm{cond}(A) = \sqrt{2}\,\|A\|\,\|A^{-1}\|. \end{aligned}$$

Corollary X.3.17 *Let A_0, A_1 be two elements of $\mathbf{GL(n)}$ with their respective QR decompositions $A_0 = Q_0R_0$ and $A_1 = Q_1R_1$. Suppose that the line segment $A(t) = (1-t)A_0 + tA_1$, $0 \leq t \leq 1$, lies in $\mathbf{GL(n)}$. Then*

$$\|Q_0 - Q_1\|_2 \leq \sqrt{2} \max_{0\leq t\leq 1} \|A(t)^{-1}\|\,\|A_0 - A_1\|_2,$$

$$\|R_0 - R_1\|_2 \leq \sqrt{2} \max_{0\leq t\leq 1} \mathrm{cond}(A(t))\|A_0 - A_1\|_2.$$

Corollary X.3.18 *Let $A_0 = Q_0R_0$ be an invertible matrix. Then for every matrix $A = QR$ close to A_0,*

$$\|Q_0 - Q\|_2 \leq \sqrt{2}\ \|A_0^{-1}\|\ \|A_0 - A\|_2 + O(\|A_0 - A\|_2^2),$$

$$\|R_0 - R\|_2 \leq \sqrt{2}\ \mathrm{cond}(A_0)\|A_0 - A\|_2 + O(\|A_0 - A\|_2^2).$$

For most other unitarily invariant norms, the norms of projections $\mathcal{P}_1$ and $\mathcal{P}_2$ onto the two summands in (X.51) are not as easy to calculate. Thus this method does not lead to attractive bounds for these norms in the case of the QR decomposition.

X.4 Appendix: Differential Calculus

We will review very quickly some basic concepts of the Fréchet differential calculus, with special emphasis on matrix analysis. No proofs are given.

Let X, Y be real Banach spaces, and let $\mathcal{L}(X, Y)$ be the space of bounded linear operators from X to Y. Let U be an open subset of X. A continuous map f from U to Y is said to be **differentiable** at a point u of U if there exists $T \in \mathcal{L}(X, Y)$ such that

$$\lim_{v \to 0} \frac{\|f(u+v) - f(u) - Tv\|}{\|v\|} = 0. \qquad \text{(X.54)}$$

It is easy to see that such a T, if it exists, is unique.

If f is differentiable at u, the operator T above is called the **derivative** of f at u. We will use for it the notation $Df(u)$. This is sometimes called the **Fréchet derivative**. If f is differentiable at every point of U, we say that it is differentiable on U.

One can see that, if f is differentiable at u, then for every $v \in X$,

$$Df(u)(v) = \frac{d}{dt}\bigg|_{t=0} f(u + tv). \qquad \text{(X.55)}$$

This is also called the **directional derivative** of f at u in the direction v.

The reader will recall from elementary calculus of functions of two variables that the existence of directional derivatives in all directions does not ensure differentiability.

Some illustrative examples are given below.

Example X.4.1 *(i) The constant function $f(x) = c$ is differentiable at all points, and $Df(x) = 0$ for all x.*

(ii) Every linear operator T is differentiable at all points, and is its own derivative; i.e., $DT(u)(v) = Tv$, for all u, v in X.

(iii) Let X, Y, Z be real Banach spaces and let $B : X \times Y \to Z$ be a bounded bilinear map. Then B is differentiable at every point, and its

derivative $DB(u,v)$ is given as

$$DB(u,v)(x,y) = B(x,v) + B(u,y).$$

(iv) Let X be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and let $f(u) = \|u\|^2 = \langle u, u \rangle$. Then f is differentiable at every point and $Df(u)(v) = 2\langle u, v \rangle$.

The next set of examples is especially important for us.

Example X.4.2 *In these examples $X = Y = \mathcal{L}(\mathcal{H})$.*
(i) Let $f(A) = A^2$. Then

$$Df(A)(B) = AB + BA.$$

(ii) More generally, let $f(A) = A^n$, $n \geq 2$. From the binomial expansion for $(A+B)^n$ one can see that

$$Df(A)(B) = \sum_{\substack{j+k=n-1 \\ j,k \geq 0}} A^j B A^k.$$

(iii) Let $f(A) = A^{-1}$ for each invertible A. Then

$$Df(A)(B) = -A^{-1} B A^{-1}.$$

(iv) Let $f(A) = A^* A$. Then

$$Df(A)(B) = A^* B + B^* A.$$

(v) Let $f(A) = e^A$. Use the formula

$$e^{A+B} - e^A = \int_0^1 e^{(1-t)A} B \, e^{t(A+B)} dt$$

(called **Dyson's expansion***) to show that*

$$Df(A)(B) = \int_0^1 e^{(1-t)A} \, B \, e^{tA} \, dt.$$

The usual rules of differentiation are valid:
If f_1, f_2 are two differentiable maps, then $f_1 + f_2$ is differentiable and

$$D(f_1 + f_2)(u) = Df_1(u) + Df_2(u).$$

The composite of two differentiable maps f and g is differentiable and we have the **chain rule**

$$D(g \cdot f)(u) = Dg(f(u)) \cdot Df(u).$$

In the special situation when g is linear, this reduces to

$$D(g \cdot f)(u) = g \cdot Df(u).$$

One important rule of differentiation for real functions is the product rule: $(fg)' = f'g + gf'$. If f and g are two maps with values in a Banach space, their product is not defined – unless the range is an algebra as well. Still, a general product rule can be established. Let f, g be two differentiable maps from X into Y_1, Y_2, respectively. Let B be a continuous bilinear map from $Y_1 \times Y_2$ into Z. Let φ be the map from X to Z defined as $\varphi(x) = B(f(x), g(x))$. Then for all u, v in X

$$D\varphi(u)(v) = B(Df(u)(v), g(u)) + B(f(u), Dg(u)(v)).$$

This is the **product rule** for differentiation. A special case of this arises when $Y_1 = Y_2 = \mathcal{L}(Y)$, the algebra of bounded operators in a Banach space Y. Now $\varphi(x) = f(x)g(x)$ is the usual product of two operators. The product rule then is

$$D\varphi(u)(v) = [Df(u)(v)] \cdot g(u) + f(u) \cdot [Dg(u)(v)].$$

Exercise X.4.3 *(i) Let f be the map $A \to A^{-1}$ on* $\mathbf{GL(n)}$. *Use the product rule to show that*

$$Df(A)(B) = -A^{-1}BA^{-1}.$$

This can also be proved directly.

(ii) Let $f(A) = A^{-2}$. Show that

$$Df(A)(B) = -A^{-1}BA^{-2} - A^{-2}BA^{-1}.$$

(iii) Obtain a formula for the derivative of the map $f(A) = A^{-n}$, $n = 3, 4, \ldots$.

Perhaps, the most useful theorem of calculus is the Mean Value Theorem.

Theorem X.4.4 *(The Mean Value Theorem) Let f be a differentiable map from an interval I of the real line into a Banach space X. Then for each closed interval $[a, b]$ contained in I,*

$$\|f(b) - f(a)\| \leq |b - a| \sup_{a \leq t \leq b} \|Df(t)\|.$$

This is the version we have used often in the book, with $[a, b] = [0, 1]$. There is a more general statement:

Theorem X.4.5 *(The Mean Value Theorem) Let f be a differentiable map from a convex subset U of a Banach space X into the Banach space Y. Let $a, b \in U$ and let L be the line segment joining them. Then*

$$\|f(b) - f(a)\| \leq \|b - a\| \sup_{u \in L} \|Df(u)\|.$$

(Note that there are three different norms that occur in the above inequality. These are the norms of the spaces Y, X, and $\mathcal{L}(X, Y)$, respectively.)

Higher order Fréchet derivatives can be identified with multilinear maps. This is explained below.

Let f be a differentiable map from X to Y. At each point u, the derivative $Df(u)$ is an element of the Banach space $\mathcal{L}(X, Y)$. Thus we have a map Df from X into $\mathcal{L}(X, Y)$, defined as $Df : u \to Df(u)$. If this map is differentiable at a point u, we say that f is **twice differentiable** at u. The derivative of the map Df at the point u is called the **second derivative** of f at u. It is denoted as $D^2 f(u)$. This is an element of the space $\mathcal{L}(X, \mathcal{L}(X, Y))$. This space is isomorphic to another Banach space, which is easier to handle.

Let $\mathcal{L}_2(X, Y)$ be the space of bounded bilinear maps from $X \times X$ into Y. The elements of this space are maps f from $X \times X$ into Y that are linear in both variables, and for whom there exists a constant c such that

$$\|f(x_1, x_2)\| \leq c\|x_1\| \, \|x_2\|$$

for all $x_1, x_2 \in X$. The infimum of all such c is called $\|f\|$. This is a norm on the space $\mathcal{L}_2(X, Y)$, and the space is a Banach space with this norm.

If φ is an element of $\mathcal{L}(X, \mathcal{L}(X, Y))$, let

$$\tilde{\varphi}(x_1, x_2) = [\varphi(x_1)](x_2) \quad \text{for } x_1, x_2 \in X.$$

Then $\tilde{\varphi} \in \mathcal{L}_2(X, Y)$. It is easy to see that the map $\varphi \to \tilde{\varphi}$ is an isometric isomorphism.

Thus the second derivative of a twice differentiable map f from X to Y can be thought of as a bilinear map from $X \times X$ to Y. It is easy to see that this map is symmetric in the two variables; i.e.,

$$D^2 f(u)(v_1, v_2) = D^2 f(u)(v_2, v_1),$$

for all u, v_1, v_2. (This symmetry property is extremely helpful in guessing the expression for the second derivative of a given map.)

Some examples on the space of matrices are given below.

Example X.4.6 *Let $X = \mathbf{M}(\mathbf{n})$ and let $f(A) = A^2, A \in \mathbf{M}(\mathbf{n})$. We have seen that $Df(A)(B) = AB + BA$ for all A, B. Note that $Df(A) = L_A + R_A$, where L_A and R_A are linear operators on $\mathbf{M}(\mathbf{n})$, the first one is the left multiplication by A and the second one is right multiplication by A. The map $Df : A \to Df(A)$ is a linear map from $\mathbf{M}(\mathbf{n})$ into $\mathcal{L}(\mathbf{M}(\mathbf{n}))$. So the derivative of this map, at each point, is the map itself. Thus for each $A, D^2 f(A) = Df$. In other words,*

$$[D^2 f(A)](B) \; = \; Df(B) \; = \; L_B + R_B.$$

If we think of $D^2 f(A)$ as a linear map from $\mathbf{M}(\mathbf{n})$ into $\mathcal{L}(\mathbf{M}(\mathbf{n}))$, we have

$$[D^2 f(A)(B_1)](B_2) \; = \; (L_{B_1} + R_{B_1})(B_2) \; = \; B_1 B_2 + B_2 B_1$$

for all B_1, B_2. *If we think of it as a bilinear map, we have*

$$[D^2 f(A)](B_1, B_2) = B_1 B_2 + B_2 B_1.$$

Note that the right-hand side is independent of A. *So the map* $A \to D^2 f(A)$ *is a constant map. These are noncommutative analogues of the facts that if* $f(x) = x^2$, *then* $f'(x) = 2x$ *and* $f''(x) = 2$.

Example X.4.7 *Let* $f(A) = A^3$. *We have seen that*

$$Df(A)(B) = A^2 B + ABA + BA^2.$$

This is the noncommutative analogue of the fact that if $f(x) = x^3$, *then* $f'(x) = 3x^2$. *What is the second derivative? From the formula* $f''(x) = 6x$, *and the fact that* $D^2 f(A)$ *is a symmetric bilinear map, we can guess that*

$$[D^2 f(A)](B_1, B_2) = AB_1B_2 + B_1AB_2 + B_1B_2A + AB_2B_1 + B_2AB_1 + B_2B_1A.$$

Prove that this indeed is the right formula for $D^2 f(A)$. *Note that the map* $A \to D^2 f(A)$ *is linear.*

Example X.4.8 *More generally, let* $f(A) = A^n$. *From the binominal theorem one can see that*

$$[D^2 f(A)](B_1, B_2) = \sum_{\substack{j+k+\ell=n-2 \\ j,k,\ell \geq 0}} [A^j B_1 A^k B_2 A^\ell + A^j B_2 A^k B_1 A^\ell].$$

Example X.4.9 *Let* $f(A) = A^{-1}, A \in \mathbf{GL(n)}$. *We know that* $Df(A)(B) = -A^{-1}BA^{-1}$, *for all* $B \in \mathbf{M(n)}$. *This is the noncommutative analogue of the formula* $(x^{-1})' = -x^{-2}$. *The analogue of the formula* $(x^{-1})'' = 2x^{-3}$ *is the following:*

$$[D^2 f(A)](B_1, B_2) = A^{-1}B_1A^{-1}B_2A^{-1} + A^{-1}B_2A^{-1}B_1A^{-1}.$$

This can be guessed from the bilinearity and symmetry properties that $D^2 f(A)$ *must have. It can be proved formally by the rules of differentiation.*

Example X.4.10 *Let* $f(A) = A^{-2}, A \in \mathbf{GL(n)}$. *We know that* $Df(A)(B) = -A^{-1}BA^{-2} - A^{-2}BA^{-1}$. *Show that*

$$\begin{aligned} [D^2 f(A)](B_1, B_2) &= A^{-2}B_1A^{-1}B_2A^{-1} + A^{-2}B_2A^{-1}B_1A^{-1} \\ &+ A^{-1}B_1A^{-2}B_2A^{-1} + A^{-1}B_2A^{-2}B_1A^{-1} \\ &+ A^{-1}B_1A^{-1}B_2A^{-2} + A^{-1}B_2A^{-1}B_1A^{-2}. \end{aligned}$$

This is the analogue of the formula $(x^{-2})'' = 6x^{-4}$.

Example X.4.11 *Let $f(A) = A^*A$. We have seen that $Df(A)(B) = A^*B + B^*A$. Show that $D^2f(A)(B_1, B_2) = B_1^*B_2 + B_2^*B_1$. Note that this expression does not involve A. So the map $A \longrightarrow D^2f(A)$ is a constant map.*

Derivatives of higher order can be defined by repeating the above procedure. The pth derivative of a map f from X to Y can be identified with a p-linear map from the space $X \times X \times \cdots \times X$ (p copies) into Y. A convenient method of calculating the pth derivative of f is provided by the formula

$$D^p f(u)(v_1, \ldots, v_p) = \frac{\partial^p}{\partial t_1 \cdots \partial t_p}\bigg|_{t_1 = \cdots = t_p = 0} f(u + t_1 v_1 + \cdots + t_p v_p). \quad \text{(X.56)}$$

Compare this with the formula (X.55) for the first derivative.

Example X.4.12 *Let $f(A) = A^n$, $A \in \mathcal{L}(\mathcal{H})$. Then for $p = 1, 2, \ldots, n$,*

$$\begin{aligned}&[D^p f(A)](B_1, \ldots, B_p) \\ &= \sum_{\sigma \in S_p} \sum_{\substack{j_i \geq 0, \\ j_1 + \cdots + j_{p+1} = n-p}} A^{j_1} B_{\sigma(1)} A^{j_2} B_{\sigma(2)} \cdots A^{j_p} B_{\sigma(p)} A^{j_{p+1}},\end{aligned}$$

where S_p is the set of all permutations on p symbols. There are $\frac{n!}{(n-p)!}$ terms in the above double sum. These are all words of length n in which $n-p$ of the letters are A and the remaining letters are $B_1, \ldots, B_p$, each occurring exactly once. Notice that this expression is linear and symmetric in each of the variables $B_1, \ldots, B_p$. When dim $\mathcal{H} = 1$, this reduces to the formula for the pth derivative of the function $f(x) = x^n$:

$$f^{(p)}(x) = n(n-1)\cdots(n-p+1)x^{n-p} = \frac{n!}{(n-p)!}x^{n-p}.$$

The reader should work out some more simple examples to see the expressions for higher derivatives.

Another important theorem of calculus, **Taylor's Theorem**, has an analogue in the Fréchet calculus. Of the different versions possible, the one that is most useful for us is given below.

Let f be a $(p+1)$-times differentiable map from a Banach space X into a Banach space Y. For $h \in X$, write $[h]^m$ to mean the m-tuple $(h, h, \ldots, h)$. Then, for all $x \in X$ and for small h,

$$\|f(x+h) - f(x) - \sum_{m=1}^{p} \frac{1}{m!} D^m f(x)([h]^m)\| = O(\|h\|^{p+1}).$$

From this we get

$$\|f(x+h) - f(x)\| \leq \sum_{m=1}^{p} \frac{1}{m!} \|D^m f(x)\| \, \|h\|^m + O(\|h\|^{p+1}).$$

Finally, let us write down formulae for higher order derivatives of the composite of two maps. These are already quite intricate for real functions. If we have $\varphi = f(g(x))$, then we have

$$\begin{aligned}
\varphi^{(1)}(x) &= f^{(1)}(g(x))g^{(1)}(x), \\
\varphi^{(2)}(x) &= f^{(2)}(g(x))[g^{(1)}(x)]^2 + f^{(1)}(g(x))g^{(2)}(x), \\
\varphi^{(3)}(x) &= f^{(3)}(g(x))[g^{(1)}(x)]^3 + 3f^{(2)}(g(x))g^{(1)}(x)g^{(2)}(x) \\
&\quad + f^{(1)}(g(x))g^{(3)}(x).
\end{aligned}$$

If X, Y, Z are Banach spaces and if f is a map from X to Y, and g a map from Y to Z, then for the derivatives of the composite map $\varphi = f \circ g$, we have the following formulae. By the chain rule,

$$D\varphi(x) = Df(g(x))Dg(x).$$

The second and the third derivatives are bilinear and trilinear maps, respectively. For them we have the formulae:

$$\begin{aligned}
[D^2\varphi(x)](x_1, x_2) &= [D^2 f(g(x))](Dg(x)(x_1), Dg(x)(x_2)) \\
&\quad + Df(g(x))([D^2 g(x)](x_1, x_2)),
\end{aligned}$$

$$\begin{aligned}
[D^3\varphi(x)](x_1, x_2, x_3) &= [D^3 f(g(x))](Dg(x)(x_1), Dg(x)(x_2), Dg(x)(x_3)) \\
&\quad + [D^2 f(g(x))](Dg(x)(x_1), [D^2 g(x)](x_2, x_3)) \\
&\quad + [D^2 f(g(x))](Dg(x)(x_2), [D^2 g(x)](x_1, x_3)) \\
&\quad + [D^2 f(g(x))](Dg(x)(x_3), [D^2 g(x)](x_1, x_2)) \\
&\quad + Df(g(x))[D^3 g(x)](x_1, x_2, x_3).
\end{aligned}$$

The reader should convince himself that considerations of domains and ranges of the maps involved, symmetry in the variables, and the demand that in the case of real functions we should recover the old formulae lead to these general formulae. He can then try proving them.

We have also used the notion of the derivative of a map between manifolds. If X and Y are differentiable manifolds in finite-dimensional vector spaces, and f is a differentiable map from X to Y, then at a point u of X the derivative $Df(x)$ is a linear map from the linear space $T_u X$ into the linear space $T_{f(u)}Y$. These are the tangent spaces to X and Y at u and $f(u)$, respectively. All manifolds we considered are subsets of $\mathbf{M}(\mathbf{n})$. Of these, $\mathbf{GL}(\mathbf{n})$, $\mathbf{P}(\mathbf{n})$, and $\mathbf{\Delta}_+(\mathbf{n})$ are open subsets of vector subspaces of $\mathbf{M}(\mathbf{n})$. So these vector spaces are the tangent spaces for the manifolds. The only closed manifold we considered is $\mathbf{U}(\mathbf{n})$. It is easy to find the tangent space at any point of this manifold. This was done in Chapter 6. Most of the results of Fréchet calculus can be restated in this setup with appropriate modifications.

X.5 Problems

Problem X.5.1. Let f be a nonnegative operator monotone function on $(0, \infty)$.

(i) Show that if A is positive and U unitary, then

$$|||f(A)U - Uf(A)||| \leq |||f(|AU - UA|)|||.$$

(ii) Let A be positive and X Hermitian. Let U be the Cayley transform of X. We have

$$\begin{aligned} U &= (X - i)(X + i)^{-1}, \\ X &= i(I + U)(I - U)^{-1} = 2i(I - U)^{-1} - iI. \end{aligned}$$

Show that

$$|||f(A)X - Xf(A)||| \leq 2\|(I - U)^{-1}\|^2 |||f(|AU - UA|)|||.$$

Use the relation between U and X again to estimate the last factor, and show that

$$|||f(A)X - Xf(A)||| \leq \frac{1 + s_1^2(X)}{2} \left|\left|\left| f\left(\frac{2}{1 + s_n^2(X)}|AX - XA|\right) \right|\right|\right|.$$

(iii) Let A, B be positive and X arbitrary. Use the above inequality to show that

$$|||f(A)X - Xf(B)||| \leq \frac{1 + s_1^2(X)}{2} \left|\left|\left| f\left(\frac{2}{1 + s_n^2(X)}|AX - XB|\right) \right|\right|\right|$$

[Hint: Use 2×2 block matrices.]

When $X = I$, this reduces to the inequality (X.5).

Problem X.5.2. Let f be a nonnegative operator monotone function. Let A, B be positive matrices and let X be any contraction. Show that

$$|||f(A)X - Xf(B)||| \leq 5/4\ |||(|AX - XB|)|||.$$

[Hint: Use the result of the preceding problem, replacing X there by $\frac{1}{2}X$.]

Problem X.5.3. From the above inequality it follows that if A, B are positive and X is any matrix, then for $0 \leq r \leq 1$,

$$\|A^r X - XB^r\| \leq 5/4\ \|X\|^{1-r} \|AX - XB\|^r.$$

Show that we have under these conditions

$$\|A^r X - XB^r\|_2 \leq \|X\|_2^{1-r} \|AX - XB\|_2^r.$$

[Hint: Reduce the general case to the special case $A = B$. Use Hölder's inequality.]

Problem X.5.4. Let A, B be any two operators. Show that

$$|||(|A| - |B|)^2||| \leq \|A + B\| \; |||A - B|||.$$

Problem X.5.5. Let A be a positive operator such that $A \geq aI \geq 0$. Show that for every X

$$2a|||X||| \leq |||AX + XA|||.$$

[Use the results in Section VII.2.]

Use this to show that if A, B are positive operators such that $A^{1/2} + B^{1/2} \geq aI \geq 0$, then

$$|||A^{1/2} - B^{1/2}||| \leq \frac{1}{a}|||A - B|||.$$

[Hint: Consider the operators $A^{1/2} + B^{1/2}$ and $A^{1/2} - B^{1/2}$.]

Problem X.5.6. Let A and B be positive operators such that $A \geq aI \geq 0$ and $B \geq bI \geq 0$. Show that for every nonnegative operator monotone function f on $(0, \infty)$

$$|||f(A) - f(B)||| \leq C(a, b)|||A - B|||,$$

where $C(a, b) = \frac{f(a)-f(b)}{a-b}$ if $a \neq b$, and $C(a, b) = f'(a)$ if $a = b$.

Problem X.5.7. Let f be a real function on $(0, \infty)$, and let $f^{(n)}$ be its nth derivative. Let f also denote the map induced by f on positive operators. Let $D^n f(A)$ be the nth order Fréchet derivative of this map at the point A. Let

$$\mathcal{D}^{(n)} = \{f : \|D^n f(A)\| = \|f^{(n)}(A)\| \text{ for all positive } A\}.$$

We have seen that every operator monotone function is in the class $\mathcal{D}^{(1)}$. Show that it is in $\mathcal{D}^{(n)}$ for all $n = 1, 2, \ldots$.

Problem X.5.8. Several examples of functions that are not operator monotone but are in $\mathcal{D}^{(1)}$ are given below.

(i) Show that for each integer n, the function $f(t) = t^n$ on $(0, \infty)$ is in the class $\mathcal{D}^{(1)}$.

(ii) Show that the function $f(t) = a_0 + a_1 t + \cdots + a_n t^n$ on $(0, \infty)$, where n is any positive integer and the coefficients a_j are nonnegative, is in the class $\mathcal{D}^{(1)}$.

(iii) Any function on $(0, \infty)$ that has a power series expansion with non-negative coefficients is in the class $\mathcal{D}^{(1)}$.

(iv) Use the Dyson expansion to show that the exponential function is in the class $\mathcal{D}^{(1)}$.

(v) Let $f(t) = \int_0^\infty e^{-\lambda t} d\mu(\lambda)$, where μ is a positive measure on $(0, \infty)$. Show that $f \in \mathcal{D}^{(1)}$. [Use part (iv).]

(vi) From the Euler's integral for the gamma function, we can write, for $r > 0$,

$$t^{-r} = \frac{1}{\Gamma(r)} \int_0^\infty e^{-\lambda t} \lambda^{r-1} d\lambda.$$

Use this to show that for each $r > 0$, the function $f(t) = t^{-r}$ on $(0, \infty)$ is in $\mathcal{D}^{(1)}$.

Problem X.5.9. The Cholesky decomposition of a positive definite matrix A is the (unique) factoring $A = R^*R$, where R is an upper triangular matrix with positive diagonal entries. This gives an invertible map Φ from the space $\mathbf{P}(\mathbf{n})$ onto the space $\boldsymbol{\Delta}_+(\mathbf{n})$. Show that

$$\|D\Phi(A)\|_2 \leq \frac{1}{\sqrt{2}} \|A\|^{1/2} \|A^{-1}\|$$

for every A. Use this to write local perturbation bounds for the map Φ.

Problem X.5.10. A matrix is called **strongly nonsingular** if all of its leading principal minors are nonzero. Such matrices form a dense open set in the space $\mathbf{M}(\mathbf{n})$. Every strongly nonsingular matrix A can be factored as $A = LR$, where L is a lower triangular matrix and R an upper triangular matrix. Further, L can be chosen to have all of its diagonal entries equal to 1. With this restriction the factoring is unique. This is the **LR decomposition** familiar in linear algebra and numerical analysis.

Let $\mathbf{S}$ be the set of strongly nonsingular matrices, Δ_1^* the set of lower triangular matrices with unit diagonal, and Δ_{ns} the set of nonsingular upper triangular matrices. Let Φ_1, Φ_2 be the maps from $\mathbf{S}$ into Δ_1^* and Δ_{ns} given by the LR decomposition.

The set $\mathbf{S}$ is an open set in $\mathbf{M}(\mathbf{n})$. So the tangent space to it at any point is $\mathbf{M}(\mathbf{n})$. The set Δ_{ns} is an open subset of the vector space Δ consisting of all upper triangular matrices. So the tangent space to Δ_{ns} at any point is Δ. The set Δ_1^* is a differentiable manifold (a Lie group, in fact). The tangent space at I to this manifold is the space Δ_0^*, consisting of lower triangular matrices with zero diagonal.

Follow the approach in Section X.3 to obtain the bounds:

$$\|D\Phi_1(A)\|_2 \leq \text{cond}(L)\|R^{-1}\|,$$

$$\|D\Phi_2(A)\|_2 \leq \text{cond}(R)\|L^{-1}\|.$$

Use these to obtain local perturbation bounds for the LR decomposition.

X.6 Notes and References

Most of the results in this chapter can be proved for infinite dimensional Hilbert space operators. Many of them are valid for operator algebras as well.

Let f be a continuous real function on an interval that contains the spectra of two Hermitian operators A, B (on a Hilbert space $\mathcal{H}$). The problem of finding bounds for $\|f(A) - f(B)\|$ in terms of $\|A - B\|$ has been investigated in great detail by many authors. Many deep results on this were obtained by the Russian school of Birman, which includes Farforovskaya, Naboko, Solomyak, and others.

When f is differentiable and f' is bounded, one would expect to find inequalities of the form

$$\|f(A) - f(B)\| \leq c\,\|f'\|_\infty\,\|A - B\|.$$

Counterexamples to show that such inequalities are not true, in general, were constructed by Yu.B. Farforovskaya, *An estimate of the norm* $\|f(B) - f(A)\|$ *for self-adjoint operators A and B*, Zap. Nauch. Sem LOMI, 56(1976) 143-162. (English translation: J. Soviet Math. 14, No. 2(1980).) It was shown by M. Sh. Birman and M.Z. Solomyak that such inequalities can be found under stronger smoothness assumptions. The reader should see their paper titled *Double Stieltjes operator integrals*, English translation, in *Topics in Mathematical Physics*, Volume 1, Consultant Bureau, New York, 1967.

Theorem X.1.1 is taken from F. Kittaneh and H. Kosaki, *Inequalities for the Schatten p-norm V*, Publ. Res. Inst. Math. Sci., 23(1987) 433-443. The inequality (X.3) was proved by M.Sh. Birman, L.S. Koplienko, and M.Z. Solomyak, *Estimates of the spectrum of the difference between fractional powers of self-adjoint operators*, Izvestiya Vysshikh Uchebnykh Zavedenni. Mat, 19 (1975) 3-10. Its generalisation in Theorem X.1.3 is due to T. Ando, *Comparison of norms* $|||f(A) - f(B)|||$ *and* $|||f(|A - B|)|||$, Math. Z., 197(1988) 403-409. Our discussion of the material between Theorem X.1.3 and Exercise X.1.7 is taken from this paper. For p-norms, the inequality (X.10) has another formulation: if A,B are positive

$$\|A^{1/t} - B^{1/t}\|_{tp} \leq \|A - B\|_p^{1/t} \quad \text{for } p \geq 1, t \geq 1.$$

The special case $t = 2$ of this was proved by R.T. Powers and E. Størmer, *Free states of the canonical anticommutation relations*, Commun. Math. Phys., 16 (1970) 1-33. The point of this formulation is that if A, B are positive Hilbert space operators and their difference $A-B$ is in the Schatten class $\mathcal{I}_p$, then $A^{1/t} - B^{1/t}$ is in the class $\mathcal{I}_{tp}$, and the above inequality is valid.

Theorem X.1.8. is due to D. Jocic and F. Kittaneh, *Some perturbation inequalities for self-adjoint operators*, J. Operator Theory, 31(1994) 3-10. The proof of Lemma X.1.9 given here is due to R. Bhatia, *A simple proof of an operator inequality of Jocic and Kittaneh*, J. Operator Theory, 31 (1994) 21-22. As in the preceding paragraph, for the Schatten p-norms, the inequality (X.12) can be written as

$$\|A - B\|_{(2m+1)p} \leq 2^{2m/2m+1} \|A^{2m+1} - B^{2m+1}\|_p^{1/2m+1},$$

for $m = 1, 2, \ldots, p \geq 1$ and Hermitian A, B. The result is valid in infinite-dimensional Hilbert spaces. A corollary of this is the statement that if the difference $A^{2m+1} - B^{2m+1}$ is in the Schatten class $\mathcal{I}_p$, then $A - B$ is in the class $\mathcal{I}_{(2m+1)p}$.

The first inequality in Problem X.5.3 was proved by G.K. Pedersen, *A commutator inequality* (unpublished note). The generalisation in Problem X.5.2, the inequalities in Problem X.5.1, and the second inequality in Problem X.5.3 are due to R. Bhatia and F. Kittaneh, *Some inequalities for norms of commutators*, SIAM J. Matrix Anal., 18(1997) to appear. The motivation for Pedersen was a result of W.B. Arveson, *Notes on extensions of C^*-algebras*, Duke Math. J., 44 (1977)329-355. Let f be a continuous function on [0,1] with $f(0) = 0$, and let $\varepsilon > 0$. Arveson showed that there exists a $\delta > 0$ such that if A and X are elements in the unit ball of a C^*-algebra and $A \geq 0$, then $\|AX - XA\| < \delta$ implies $\|f(A)X - Xf(A)\| < \varepsilon$. The inequality in Problem X.5.3 is a quantitative version of this for the special class of functions $f(t) = t^r, 0 \leq r \leq 1$. Weaker results proved earlier and their applications may be found in C.L. Olsen and G.K. Pedersen, *Corona C^* - algebras and their applications to lifting problems*, Math. Scand., 64(1989) 63-86. It is conjectured that the factor 5/4 occurring in these inequalities can be replaced by 1.

The inequality (X.20) for $p = 1$ was proved by H. Kosaki, *On the continuity of the map $\varphi \longrightarrow |\varphi|$ from the predual of a W^*-algebra*, J. Funct. Anal., 59(1984) 123-131. For the Schatten p-norms, $p \geq 2$, the inequality (X.18) was proved by F. Kittaneh and H. Kosaki in their paper cited above. The other parts of Theorems X.2.3 and X.2.4, and Theorem X.2.1 were proved by R. Bhatia, *Perturbation inequalities for the absolute value map in norm ideals of operators*, J. Operator Theory, 19(1988) 129-136.

The constant $\sqrt{2n}$ in (X.25) can be replaced by a factor $c_n \approx \log n$. This has been known for some time and is related to other important problems in operator theory. See two papers by A. McIntosh, *Counterexample to a question on commutators*, Proc, Amer. Math. Soc., 29(1971) 337-340, and

Functions and derivations of C^*-algebras, J. Funct. Anal., 30(1978)264-275. It is also known that such a factor is indeed necessary, both for the operator norm and for the corresponding inequality for the trace norm $\|.\|_1$. This implies that, if $\mathcal{H}$ is infinite-dimensional, then the map $A \mapsto |A|$ on $\mathcal{L}(\mathcal{H})$ is not Lipschitz continuous. Nor is it Lipschitz continuous on the Schatten ideal $\mathcal{I}_1$. The inequality (X.24) due to Araki and Yamagami, on the other hand, shows that on the Hilbert-Schmidt ideal $\mathcal{I}_2$ this map is continuous. For other values of $p, 1 < p < \infty$, E.B. Davies, *Lipschitz continuity of functions of operators in the Schatten classes*, J. London Math. Soc., 37(1988)148-157, showed that there exists a constant γ_p that depends on p, but not on the dimension n, such that

$$\| \, |A| - |B| \, \|_p \leq \gamma_p \, \|A - B\|,$$

for all A, B. Theorem X.2.5 was proved in T. Kato, *Continuity of the map $S \to |S|$ for linear operators*, Proc. Japan Acad. 49 (1973) 157-160, and interpreted to mean that the map $A \mapsto |A|$ is "almost Lipschitz". Results close to this were obtained by Yu.B. Farforovskaya in the papers cited above.

Bounds like the ones in Section X.3 have been of interest to numerical analysts and physicists. References to much of this work may be found in R. Bhatia, *Matrix factorizations and their perturbations*, Linear Algebra Appl., 197/198 (1994) 245-276. Theorem X.3.1, and the proof given here, are due to C.J. Kenney and A.J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Analysis, 10(1989) 191-209. Theorem X.3.4 was proved in R. Bhatia, *First and second order perturbation bounds for the operator absolute value*, Linear Algebra Appl., 208/209 (1994) 367-376. Theorems X.3.3, X.3.6, X.3.7, and X.3.8 are also proved in this paper. The inequality in Problem X.5.5 is taken from J.L. van Hemmen and T. Ando, *An inequality for trace ideals*, Commun. Math. Phys., 76(1980) 143-148. This paper has references to physics literature, where such inequalities are used. The inequality in Problem X.5.6 is proved in the paper by F. Kittaneh and H. Kosaki cited earlier. Most of the results after Theorem X.3.9 in Section X.3 were proved by R. Bhatia and K. Mukherjea, *Variation of the unitary part of a matrix*, SIAM J. Matrix Analysis, 15(1994) 1007-1014. The full potential of this method was exploited in the paper cited at the beginning of this paragraph, where several other matrix decompositions of interest in numerical analysis are studied. The results of Problem X.5.9 and X.5.10 are obtained in this paper. (Some of these were proved earlier using different methods by A. Barrlund, R. Mathias, G.W. Stewart, and J. G. Sun.)

Bounds for the second derivative of the map $A \to |A|$ are obtained in R. Bhatia, *First and second order perturbation bounds for the operator absolute value*, Linear Algebra Appl., 208/209 (1994) 367-376; and for derivatives of higher orders in R. Bhatia, *Perturbation bounds for the operator absolute value*, Linear Algebra Appl., 226(1995) 639-645. The reader may try to

prove such inequalities using the methods explained in Section X.4. Since this map is the composite of two maps, $A \to A^*A \to (A^*A)^{1/2}$, its analysis can be broken into two parts. No good bounds of higher order are known for other matrix decompositions.

Results in Parts (v) and (vi) of Problem X.5.8 are taken from R. Bhatia and K.B. Sinha, *Variation of real powers of positive operators*, Indiana Univ. Math. J., 43(1994)913-925. In this paper it is also shown that the functions $f(t) = t^r$ on $(0, \infty)$ belong to the class $\mathcal{D}^{(1)}$ if $r \geq 2$, but not if $1 < r < \sqrt{2}$. We have already seen that these functions are in $\mathcal{D}^{(1)}$ for all real numbers $r \leq 1$.

In Section X.4 we have given a bare outline of differential calculus. More on this may be found in J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, 1960, and in A. Ambrosetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge University Press, 1993. For calculus on manifolds, the reader could see S. Lang, *Introduction to Differentiable Manifolds*, John Wiley, 1962. In our exposition we have included several examples of matrix functions and formulae for higher derivatives of composite maps that are not easily found in other sources.

References

C. Akemann, J. Anderson and G. Pedersen, *Triangle inequalities in operator algebras*, Linear and Multilinear Algebra, 11(1982) 167-178.

A. Ambrosetti and G. Prodi, *A Primer of Nonlinear Analysis*, Cambridge University Press, 1993.

A.R. Amir-Moez, *Extreme Properties of Linear Transformations and Geometry in Unitary Spaces*, Texas Tech. University, 1968.

W.N. Anderson and G.E. Trapp, *A class of monotone operator functions related to electrical network theory*, Linear Algebra Appl., 15(1975) 53-67.

T. Ando, *Topics on operator inequalities*, Hokkaido University, Sapporo, 1978.

T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26(1979) 203-241.

T. Ando, *Inequalities for permanents*, Hokkaido Math. J., 10(1981) 18-36.

T. Ando, *Comparison of norms* $|||f(A)-f(B)|||$ *and* $|||f(|A-B|)|||$, Math. Z., 197(1988) 403-409.

T. Ando, *Majorization, doubly stochastic matrices and comparison of eigenvalues*, Linear Algebra Appl., 118(1989) 163-248.

T. Ando, *Matrix Young inequalities*, Operator Theory: Advances and Applications, 75(1995) 33-38.

T. Ando, *Bounds for the antidistance*, J. Convex Analysis, 2(1996) 1-3.

T. Ando and R. Bhatia, *Eigenvalue inequalities associated with the Cartesian decomposition*, Linear and Multilinear Algebra, 22 (1987) 133-147.

T. Ando and F. Hiai, *Log majorization and complementary Golden- Thompson type inequalities*, Linear Algebra Appl., 197/198(1994) 113-131.

H. Araki, *On an inequality of Lieb and Thirring*, Letters in Math. Phys., 19(1990) 167-170.

H. Araki and S. Yamagami, *An inequality for the Hilbert-Schmidt norm*, Commun. Math. Phys., 81(1981) 89-98.

N. Aronszajn, *Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues. I. Operators in a Hilbert space*, Proc. Nat. Acad. Sci. U.S.A., 34(1948) 474-480.

W.B. Arveson, *Notes on extensions of C^*-algebras*, Duke Math. J., 44 (1977) 329-355.

J.S. Aujla and H.L. Vasudeva, *Convex and monotone operator functions*, Ann. Polonici Math., 62(1995) 1-11.

F.L. Bauer and C.T. Fike, *Norms and exclusion theorems*, Numer. Math. 2(1960) 137-141.

H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser, 1984.

N. Bebiano, *New developments on the Marcus-Oliviera conjecture*, Linear Algebra Appl., 197/198(1994) 793-802.

G.R. Belitskii and Y.I. Lyubich, *Matrix Norms and Their Applications*, Birkhäuser, 1988.

J. Bendat and S. Sherman, *Monotone and convex operator functions*, Trans. Amer. Math. Soc., 79(1955) 58-71.

F. Berezin and I.M. Gel'fand, *Some remarks on the theory of spherical functions on symmetric Riemannian manifolds*, Trudi Moscow Math. Ob., 5(1956) 311-351.

R. Bhatia, *On the rate of change of spectra of operators II*, Linear Algebra Appl., 36(1981) 25-32.

R. Bhatia, *Analysis of spectral variation and some inequalities*, Trans. Amer. Math. Soc., 272(1982) 323-332.

R, Bhatia, *Some inequalities for norm ideals*, Commun. Math. Phys., 111(1987) 33-39.

R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman, 1987.

R. Bhatia, *Perturbation inequalities for the absolute value map in norm ideals of operators*, J. Operator Theory, 19(1988) 129-136.

R. Bhatia, *On residual bounds for eigenvalues*, Indian J. Pure Appl. Math., 23(1992) 865-866.

R. Bhatia, *A simple proof of an operator inequality of Jocic and Kittaneh*, J. Operator Theory, 31(1994) 21-22.

R. Bhatia, *Matrix factorizations and their perturbations*, Linear Algebra Appl., 197/198(1994) 245-276.

R. Bhatia, *First and second order perturbation bounds for the operator absolute value*, Linear Algebra Appl., 208/209(1994) 367-376.

R. Bhatia, *Perturbation bounds for the operator absolute value*, Linear Algebra Appl., 226(1995) 639-645.

R. Bhatia and C. Davis, *A bound for the spectral variation of a unitary operator*, Linear and Multilinear Algebra, 15(1984) 71-76.

R. Bhatia and C. Davis, *Concavity of certain functions of matrices*, Linear and Multilinear Algebra, 17(1985) 155-164.

R. Bhatia and C. Davis, *More matrix forms of the arithmetic-geometric mean inequality*, SIAM J. Matrix Analysis, 14(1993) 132-136.

R. Bhatia and C. Davis, *Relations of linking and duality between symmetric gauge functions*, Operator Theory: Advances and Applications, 73(1994) 127-137.

R. Bhatia and C. Davis, *A Cauchy-Schwarz inequality for operators with applications*, Linear Algebra Appl., 223(1995) 119-129.

R. Bhatia, C. Davis, and F. Kittaneh, *Some inequalities for commutators and an application to spectral variation*, Aequationes Math., 41(1991) 70-78.

R. Bhatia, C. Davis and A. McIntosh, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52/53(1983) 45-67.

R. Bhatia and L. Elsner, *On the variation of permanents*, Linear and Multilinear Algebra, 27(1990) 105-110.

R. Bhatia and L. Elsner, *Symmetries and variation of spectra*, Canadian J. Math., 44(1992) 1155-1166

R. Bhatia and L. Elsner, *The q-binomial theorem and spectral symmetry*, Indag. Math., N.S., 4(1993) 11-16.

R. Bhatia, L. Elsner, and G. Krause, *Bounds for the variation of the roots of a polynomial and the eigenvalues of a matrix*, Linear Algebra Appl., 142(1990) 195-209.

R. Bhatia, L. Elsner, and G. Krause, *Spectral variation bounds for diagonalisable matrices*, Preprint 94-098, SFB 343, University of Bielefeld, Aequationes Math., to appear.

R. Bhatia and S. Friedland, *Variation of Grassmann powers and spectra*, Linear Algebra Appl., 40(1981) 1-18.

R. Bhatia and J.A.R. Holbrook, *Short normal paths and spectral variation*, Proc. Amer. Math. Soc., 94(1985) 377-382.

R. Bhatia and J.A.R. Holbrook, *Unitary invariance and spectral variation*, Linear Algebra Appl., 95(1987) 43-68.

R. Bhatia and J.A.R. Holbrook, *A softer, stronger Lidskii theorem*, Proc. Indian Acad. Sci. (Math. Sci.), 99(1989) 75-83.

R. Bhatia, R. Horn, and F. Kittaneh, *Normal approximants to binormal operators*, Linear Algebra Appl., 147(1991) 169-179.

R. Bhatia and F. Kittaneh, *On some perturbation inequalities for operators*, Linear Algebra Appl., 106(1988) 271-279.

R. Bhatia and F. Kittaneh, *On the singular values of a product of operators*, SIAM J. Matrix Analysis, 11(1990) 272-277.

R. Bhatia and F. Kittaneh, *Some inequalities for norms of commutators*, SIAM J. Matrix Analysis, 18(1997) to appear.

R. Bhatia, F. Kittaneh and R.-C. Li, *Some inequalities for commutators and an application to spectral variation II*, Linear and Multilinear Algebra, to appear.

R. Bhatia and K.K. Mukherjea, *On the rate of change of spectra of operators*, Linear Algebra Appl., 27(1979) 147-157.

R. Bhatia and K.K. Mukherjea, *The space of unordered types of complex numbers*, Linear Algebra Appl., 52/53(1983) 765-768.

R. Bhatia and K. Mukherjea, *Variation of the unitary part of a matrix*, SIAM J. Matrix Analysis, 15(1994) 1007-1014.

R. Bhatia and P. Rosenthal, *How and why to solve the equation $AX-XB=Y$*, Bull. London Math. Soc., 29(1997) to appear.

R. Bhatia and K.B. Sinha, *Variation of real powers of positive operators*, Indiana Univ. Math. J., 43(1994) 913-925.

M.Sh. Birman, L.S. Koplienko, and M.Z. Solomyak, *Estimates of the spectrum of the difference between fractional powers of self-adjoint operators*, Izvestiya Vysshikh Uchebnykh Zavedenni. Mat, 19(1975) 3-10.

M. Sh. Birman and M.Z. Solomyak *Double Stieltjes operator integrals*, English translation, in *Topics in Mathematical Physics*, Volume 1, Consultant Bureau, New York, 1967.

A. Björck and G.H. Golub, Numerical methods for computing angles between linear subspaces, Math. Comp. 27(1973) 579-594.

R. Bouldin, *Best approximation of a normal operator in the Schatten p-norm*, Proc. Amer. Math. Soc., 80(1980) 277-282.

A.L. Cauchy, *Sur l'équation á l'aide de laquelle on détermine les inégalités séculaires des mouvements des planétes*,1829, Oeuvres Complétes, (IInd Série) Volume 9, Gauthier-Villars.

F. Chatelin, Spectral Approximation of Linear Operators, Academic Press 1983.

M.D. Choi, *Almost commuting matrices need not be nearly commuting*, Proc. Amer. Math. Soc. 102(1988) 529-533.

J.E. Cohen, *Inequalities for matrix exponentials*, Linear Algebra Appl., 111(1988) 25-28.

J.E. Cohen, S. Friedland, T. Kato, and F.P. Kelly, *Eigenvalue inequalities for products of matrix exponentials*, Linear Algebra Appl., 45(1982) 55-95.

A. Connes and E. Stφrmer, *Entropy for automorphisms of II_1 von Neumann algebras*, Acta Math., 134(1975) 289-306.

H.O. Cordes, *Spectral Theory of Linear Differential Operators and Comparison Algebras*, Cambridge University Press, 1987.

R. Courant, *Über die Eigenwerte bei den Differentialgleichungen der mathematischen Physik*, Math. Z., 7(1920) 1-57.

R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Wiley, 1953.

Ju. L. Daleckii and S.G. Krein, *Formulas of differentiation according to a parameter of functions of Hermitian operators*, Dokl. Akad. Nauk SSSR, 76(1951) 13-16.

E.B. Davies, *Lipschitz continuity of functions of operators in the Schatten classes*, J. London Math. Soc., 37(1988)148-157.

C. Davis, *Separation of two linear subspaces*, Acta Sci. Math. (Szeged), 19(1958) 172-187.

C. Davis, *Notions generalizing convexity for functions defined on spaces of matrices*, in *Convexity: Proceedings of Symposia in Pure Mathematics*, American Mathematical Society, (1963) 187-201.

C. Davis, *The rotation of eigenvectors by a perturbation*, J. Math. Anal. Appl., 6(1963) 159-173.

C. Davis, *The Toeplitz-Hausdorff theorem explained*, Canad. Math. Bull., 14(1971) 245-246.

C. Davis and W.M. Kahan, *The rotation of eigenvectors by a perturbation III*, SIAM J. Numer. Anal. 7(1970) 1-46.

J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, 1960.

W.F. Donoghue, *Monotone Matrix Functions and Analytic Continuation*, Springer-Verlag , 1974.

L. Elsner, *On the variation of the spectra of matrices*, Linear Algebra Appl., 47(1982) 127-138.

L. Elsner, *An optimal bound for the spectral variation of two matrices*, Linear Algebra Appl., 71(1985) 77-80.

L. Elsner and S. Friedland, *Singular values, doubly stochastic matrices and applications*, Linear Algebra Appl., 220(1995) 161-169.

L. Elsner and C. He, *Perturbation and interlace theorems for the unitary eigenvalue problem*, Linear Algebra Appl., 188/189(1993) 207-229.

L. Elsner, C. Johnson, J. Ross and J. Schönheim, *On a generalised matching problem arising in estimating the eigenvalue variation of two matrices*, European J. Combinatorics, 4(1983) 133-136.

L. Elsner and M.H.C. Paardekooper, *On measures of nonnormality of matrices*, Linear Algebra Appl., 92(1987) 107-124.

Ky Fan, *On a theorem of Weyl concerning eigenvalues of linear transformations I*, Proc. Nat. Acad. Sci., U.S.A., 35(1949) 652-655.

Ky Fan, *On a theorem of Weyl concerning eigenvalues of linear transformations II*, Proc. Nat. Acad. Sci., U.S.A., 36(1950) 31-35.

Ky Fan, *A minimum property of the eigenvalues of a Hermitian transformation*, Amer. Math. Monthly, 60(1953) 48-50.

Ky Fan and A.J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6(1955) 111-116.

Yu.B. Farforovskaya, *An estimate of the norm* $\|f(B) - f(A)\|$ *for selfadjoint operators* A *and* B, Zap. Nauch. Sem LOMI, 56(1976) 143-162. (English translation: J. Soviet Math. 14, No. 2(1980).)

M. Fiedler, *Bounds for the determinant of the sum of Hermitian matrices*, Proc. Amer. Math. Soc., 30(1971) 27-31.

E. Fischer, *Über Quadratische Formen mit reellen Koeffizienten*, Monatsh. Math. Phys., 16(1905) 234-249.

C.-K. Fong and J.A.R. Holbrook, *Unitarily invariant operator norms*, Canad. J. Math., 35(1983) 274-299.

C.-K. Fong, H. Radjavi and P. Rosenthal, *Norms for matrices and operators*, J. Operator Theory, 18(1987) 99-113.

M. Fujii and T. Furuta, *Löwner-Heinz, Cordes and Heinz-Kato inequalities*, Math. Japonica, 38(1993) 73-78.

T. Furuta, *Norm inequalities equivalent to Löwner-Heinz theorem*, Reviews in Math. Phys., 1(1989) 135-137.

F.R. Gantmacher, *Matrix Theory,* 2 volumes, Chelsea, 1959.

L. Garding, *Linear hyperbolic partial differential equations with constant coefficients*, Acta Math., 84(1951) 1-62.

L. Garding, *An inequality for hyperbolic polynomials*, J. Math. Mech., 8(1959) 957-966.

I.M. Gel'fand and M. Naimark, *The relation between the unitary representations of the complex unimodular group and its unitary subgroup*, Izv Akad. Nauk SSSR Ser. Mat. 14(1950) 239-260.

S.A. Gersgorin, *Über die Abrenzung der Eigenwerte einer Matrix*, Izv. Akad. Nauk SSSR, Ser. Fiz. - Mat., 6(1931) 749-754.

I.C. Gohberg and M.G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, American Math. Society, 1969.

S. Golden, *Lower bounds for the Helmholtz function*, Phys. Rev. B, 137(1965) 1127-1128.

J.A. Goldstein and M. Levy, *Linear algebra and quantum chemistry*, American Math. Monthly, 78(1991) 710-718.

G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 2nd ed., 1989.

W. Greub, *Multilinear Algebra*, 2nd ed., Springer-Verlag, 1978.

P.R. Halmos, *Finite-dimensional Vector Spaces*, Van Nostrand, 1958.

P.R. Halmos, *Positive approximants of operators*, Indiana Univ. Math. J. 21(1972) 951-960.

P.R. Halmos, *Spectral approximants of normal operators*, Proc. Edinburgh Math. Soc., 19(1974) 51-58.

P.R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, 1982.

F. Hansen and G.K. Pederesen, *Jensen's inequality for operators and Löwner's theorem*, Math. Ann., 258(1982) 229-241.

G.H. Hardy, J.E. Littlewood and G. Polya, *Inequalities*, Cambridge University Press, 1934.

E. Heinz, *Beiträge zur Störungstheoric der Spektralzerlegung*, Math. Ann., 123(1951) 415-438.

R.H. Herman and A. Ocneanu, *Spectral analysis for automorphisms of UHF C^*-algebras*, J. Funct. Anal., 66(1986) 1-10.

P. Henrici, *Bounds for iterates, inverses, spectral variation and fields of values of nonnormal matrices*, Numer. Math., 4(1962) 24-39.

F. Hiai and Y. Nakamura, *Majorisation for generalised s-numbers in semifinite von Neumann algebras*, Math. Z., 195(1987) 17-27.

F. Hiai and D. Petz, *The Golden-Thompson trace inequality is complemented*, Linear Algebra Appl., 181(1993) 153-185.

N.J. Higham, *Matrix nearness problems and applications*, in *Applications of Matrix Theory*, Oxford University Press, 1989.

A.J. Hoffman and H.W. Wielandt, *The variation of the spectrum of a normal matrix*, Duke Math J., 20(1953) 37-39.

K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed., Prentice Hall, 1971.

J.A.R. Holbrook, *Spectral variation of normal matrices*, Linear Algebra Appl., 174(1992) 131-144.

A. Horn, *Eigenvalues of sums of Hermitian matrices*, Pacific J. Math., 12(1962) 225-242.

A. Horn and R. Steinberg, *Eigenvalues of the unitary part of a matrix*, Pacific J. Math., 9(1959) 541-550.

R.A. Horn, *The Hadamard product*, in C.R. Johnson, ed., *Matrix Theory and Applications*, American Mathematical Society, 1990.

R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1990.

R.A. Horn and R. Mathias, *An analog of the Cauchy-Schwarz inequality for Hadmard products and unitarily invariant norms*, SIAM J. Matrix Analysis, 11(1990) 481-498.

R.A. Horn and R. Mathias, *Cauchy-Schwarz inequalities associated with positive semidefinite matrices*, Linear Algebra Appl., 142(1990) 63-82.

N.M. Hugenholtz and R.V. Kadison, *Automorphisms and quasi-free states of the CAR algebra*, Commun. Math. Phys. 43(1975) 181-197.

C. Jordan, *Essai sur la géométrie à n dimensions*, Bull. Soc. Math. France, 3 (1875) 103-174.

W. Kahan, *Numerical linear algebra*, Canadian Math. Bull., 9(1966) 757-801.

W. Kahan, *Inclusion theorems for clusters of eigenvalues of Hermitian matrices*, Technical Report, Computer Science Department, University of Toronto, 1967.

W. Kahan, *Every $n \times n$ matrix Z with real spectrum satisfies* $\|Z-Z^*\| \leq \|Z+Z^*\|(\log_2 n+0.038)$, Proc. Amer. Math. Soc., 39(1973) 235-241.

W. Kahan, *Spectra of nearly Hermitian matrices*, Proc. Amer. Math. Soc., 48(1975) 11-17.

W. Kahan, B.N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal. 19(1982) 470-484.

T. Kato, *Notes on some inequalities for linear operators*, Math. Ann., 125(1952) 208-212.

T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, 1966.

T. Kato, *Continuity of the map $S \to |S|$ for linear operators*, Proc. Japan Acad. 49(1973) 157-160.

C.J. Kenney and A.J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Analysis, 10(1989) 191-209.

F. Kittaneh, *On Lipschitz functions of normal operators*, Proc. Amer. Math. Soc., 94(1985) 416-418.

F. Kittaneh, *Inequalities for the Schatten p-norm IV*, Commun. Math. Phys., 106(1986) 581-585.

F. Kittaneh, *A note on the arithmetic-geometric mean inequality for matrices*, Linear Algebra Appl., 171(1992) 1-8.

F. Kittaneh and D. Jocic, *Some perturbation inequalities for self-adjoint operators*, J. Operator Theory, 31(1994) 3-10.

F. Kittaneh and H. Kosaki, *Inequalities for the Schatten p-norm V*, Publ. Res. Inst. Math. Sci., 23(1987) 433-443.

A. Korányi, *On a theorem of Löwner and its connections with resolvents of selfadjoint transformations*, Acta Sci. Math. Szeged, 17(1956) 63-70.

H. Kosaki, *On the continuity of the map $\varphi \to |\varphi|$ from the predual of a W^*-algebra*, J. Funct. Anal., 59(1984) 123-131.

B. Kostant, *On convexity, the Weyl group and the Iwasawa decomposition*, Ann. Sci. E.N.S., 6(1973) 413-455.

F. Kraus, *Über konvexe Matrixfunktionen*, Math. Z., 41(1936) 18-42.

G. Krause, *Bounds for the variation of matrix eigenvalues and polynomial roots*, Linear Algebra Appl., 208/209(1994) 73-82.

F. Kubo and T. Ando, *Means of positive linear operators*, Math. Ann., 249(1980) 205-224.

S. Lang, *Introduction to Differentiable Manifolds*, John Wiley, 1962.

P. D. Lax, *Differential equations, difference equations and matrix theory*, Comm. Pure Appl. Math., 11(1958) 175-194.

A. Lenard, *Generalization of the Golden-Thompson inequality*, Indiana Univ. Math. J. 21(1971) 457-468.

C.-K. Li, *Some aspects of the theory of norms*, Linear Algebra Appl., 212/213(1994) 71-100.

C.-K. Li and N.-K. Tsing, *On the unitarily invariant norms and some related results*, Linear and Multilinear Algebra, 20(1987) 107-119.

C.-K. Li and N.-K. Tsing, *Norms that are invariant under unitary similarities and the C-numerical radii*, Linear and Multilinear Algebra, 24(1989) 209-222.

R.-C. Li, *New perturbation bounds for the unitary polar factor*, SIAM J. Matrix Analysis., 16(1995) 327-332.

R.-C.Li, *Norms of certain matrices with applications to variations of the spectra of matrices and matrix pencils*, Linear Algebra Appl., 182(1993) 199-234.

B.V. Lidskii, *Spectral polyhedron of a sum of two Hermitian matrices*, Functional Analysis and Appl., 10(1982) 76-77.

V.B. Lidskii, *On the proper values of a sum and product of symmetric matrices*, Dokl. Akad. Nauk SSSR, 75(1950) 769-772.

E.H. Lieb, *Convex trace functions and the Wigner-Yanase-Dyson conjecture*, Advances in Math., 11(1973) 267-288.

E.H. Lieb, *Inequalities for some operator and matrix functions*, Advances in Math., 20(1976) 174-178.

H. Lin, *Almost commuting selfadjoint matrices and applications*, preprint, 1995.

G. Lindblad, *Entropy, information and quantum measurements*, Commun. Math. Phys., 33(1973) 305-322.

J.H. van Lint, *The van der Waerden conjecture: two proofs in one year*, Math., Intelligencer, 4(1982) 72-77.

P.O. Löwdin, *On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals*, J. Chem. Phys., 18(1950) 365-374.

K. Löwner, *Über monotone Matrixfunctionen*, Math. Z., 38(1934) 177-216.

T.-X. Lu, *Perturbation bounds for eigenvalues of symmetrizable matrices*, Numerical Mathematics: a Journal of Chinese Universities, 16(1994) 177-185 (in Chinese).

G. Lumer and M. Rosenblum, *Linear operator equations*, Proc. Amer. Math. Soc., 10(1959) 32-41.

M. Marcus, *Finite-dimensional Multilinear Algebra*, 2 volumes, Marcel Dekker, 1973 and 1975.

M. Marcus and L. Lopes, *Inequalities for symmetric functions and Hermitian matrices*, Canad. J. Math., 9(1957) 305-312.

M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Prindle, Weber and Schmidt, 1964, reprinted by Dover in 1992.

M. Marcus and M. Newmann, *Inequalities for the permanent function*, Ann. of Math., 75(1962) 47-62.

A.S. Markus, *The eigen-and singular values of the sum and product of linear operators*, Russian Math. Surveys, 19(1964) 92-120.

A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, 1979.

R. Mathias, *The Hadamard operator norm of a circulant and applications*, SIAM J. Matrix Analysis, 14(1993) 1152-1167.

R. McEachin, *A sharp estimate in an operator inequality*, Proc. Amer. Math. Soc., 115(1992) 161-165.

R. McEachin, *Analyzing specific cases of an operator inequality*, Linear Algebra Appl., 208/209(1994) 343-365.

A. McIntosh, *Counterexample to a question on commutators*, Proc, Amer. Math. Soc., 29(1971) 337-340.

A. McIntosh, *Functions and derivations of C^*-algebras*, J. Funct. Anal., 30(1978)264-275.

A. McIntosh, *Heinz inequalities and perturbation of spectral families*, Macquarie Math. Reports, 1979.

M.L. Mehta, *Matrix Theory*, 2nd ed., Hindustan Publishing Co., 1989.

R. Merris and J.A. Dias da Silva, *Generalized Schur functions*, J. Algebra, 35(1975) 442-448.

H. Minc, *Permanents*, Addison Wesley, 1978.

B.A. Mirman, *Numerical range and norm of a linear operator*, Trudy Seminara po Funkcional' nomu Analizu, No. 10 (1968), 51-55.

L. Mirsky, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc., 33(1958) 14-21.

L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math., Oxford Ser. (2), 11(1960) 50-59.

Y. Nakamura, *Numerical range and norm*, Math. Japonica, 27(1982) 149-150.

A. Nudel'man and P. Svarcman, *The spectrum of a product of unitary matrices*, Uspehi Mat. Nauk 13(1958) 111-117.

M. Ohya and D. Petz, *Quantum Entropy and Its use*, Springer-Verlag, 1993.

K. Okubo, *Hölder-type norm inequalities for Schur products of matrices*, Linear Algebra Appl., 91(1987) 13-28.

C.L. Olsen and G.K. Pedersen, *Corona C^* - algebras and their applications to lifting problems*, Math. Scand., 64(1989) 63-86.

M. Omladic and P. Semrl, *On the distance between normal matrices*, Proc. Amer. Math. Soc., 110(1990) 591-596.

A. Ostrowski, *Recherches sur ℓa méthode de Gräffe et ℓes zeros des polynômes et des series de Laurent*, Acta Math., 72(1940) 99-257.

A. Ostrowski, *Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen*, Jber. Deut. Mat. - Verein, 60(1957) 40-42.

A. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, 1960.

C.C. Paige and M. Wei, *History and generality of the CS Decomposition*, Linear Algebra Appl., 208/209(1994) 303-326.

B.N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, 1980.

K.R. Parthasarathy, *Eigenvalues of matrix-valued analytic maps*, J. Austral. Math. Soc.(Ser. A), 26(1978) 179-197.

C. Pearcy and A. Shields, *Almost commuting matrices*, J. Funct. Anal., 33(1979) 332-338.

G.K. Pedersen, *A commutator inequality* (unpublished note).

D. Petz, *Quasi-entropies for finite quantum systems*, Rep. Math. Phys., 21(1986) 57-65.

D. Petz, *A survey of certain trace inequalities*, Banach Centre Publications 30(1994) 287-298.

D. Phillips, *Improving spectral variation bounds with Chebyshev polynomials*, Linear Algebra Appl., 133(1990) 165-173.

J. Phillips, *Nearest normal approximation for certain normal operators*, Proc. Amer. Math. Soc., 67(1977) 236-240.

A. Pokrzywa, *Spectra of operators with fixed imaginary parts*, Proc. Amer. Math. Soc., 81(1981) 359-364.

I.R. Porteous, *Topological Geometry*, Cambridge University Press, 1981.

R.T. Powers and E. Stϕrmer, *Free states of the canonical anticommutation relations*, Commun. Math. Phys., 16(1970) 1-33.

J.F. Queiró and A.L. Duarte, *On the Cartesian decomposition of a matrix*, Linear and Multilinear Algebra, 18(1985) 77-85.

Lord Rayleigh, *The Theory of Sound*, reprinted by Dover, 1945.

M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Volume 4, Academic Press, 1978.

R.C. Riddell, *Minimax problems on Grassmann manifolds*, Advances in Math., 54(1984) 107-199.

M. Rosenblum, *On the operator equation* $BX - XA = Q$, Duke Math. J., 23(1956) 263-270.

S. Ju. Rotfel'd, *The singular values of a sum of completely continuous operators*, Topics in Mathematical Physics, Consultants Bureau, 1969, Vol. 3, 73-78.

A. Ruhe, *Closest normal matrix finally found!* BIT, 27(1987) 585-598.

D. Ruelle, *Statistical Mechanics*, Benjamin, 1969.

R. Schatten, *Norm Ideals of Completely Continuous Operators*, Springer-Verlag , 1960.

A. Schönhage, *Quasi-GCD computations*, J. Complexity, 1(1985) 118-137.

J.P. Serre, *Linear Representations of Finite Groups*, Springer-Verlag, 1977.

H.S. Shapiro, *Topics in Approximation Theory*, Springer Lecture Notes in Mathematics, Vol. 187, 1971.

B. Simon, *Trace Ideals and Their Applications*, Cambridge University Press, 1979.

S. Smale, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (New Series), 4(1981) 1-36.

M.F. Smiley, *Inequalities related to Lidskii's*, Proc. Amer. Math. Soc., 19(1968) 1029-1034.

G. Sparr, *A new proof of Löwner's theorem on monotone matrix functions*, Math. Scand., 47(1980) 266-274.

G.W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15(1973) 727-764.

G.W. Stewart, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Rev., 19(1977) 634-662.

G.W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, 1990.

J.-G. Sun, *On the perturbation of the eigenvalues of a normal matrix*, Math. Numer. Sinica, 6(1984) 334-336.

J.-G. Sun, *On the variation of the spectrum of a normal matrix*, Linear Algebra Appl., 246(1996) 215-222.

V.S. Sunder, *Distance between normal operators*, Proc. Amer. Math. Soc., 84(1982) 483-484.

V.S. Sunder, *On permutations, convex hulls and normal operators*, Linear Algebra Appl., 48(1982) 403-411.

J.H. Sylvester, *Sur l'equation en matrices* $px = xq$, C.R. Acad. Sci. Paris, 99(1884) 67-71 and 115-116.

B. Sz.-Nagy, *Über die Ungleichung von H. Bohr*, Math. Nachr., 9(1953) 255-259.

P. Tarazaga, *Eigenvalue estimates for symmetric matrices*, Linear Algebra Appl., 135(1990) 171-179.

C.J. Thompson, *Inequality with applications in statistical mechanics*, J. Math. Phys., 6(1965) 1812-1813.

C.J. Thompson, *Inequalities and partial orders on matrix spaces*, Indiana Univ. Math. J. 21(1971) 469-480.

R.C. Thompson, *Principal submatrices II*, Linear Algebra Appl., 1(1968) 211-243.

R.C. Thompson, *Principal submatrices IX*, Linear Algebra Appl., 5(1972) 1-12.

R.C. Thompson, *On the eigenvalues of a product of unitary matrices*, Linear and Multilinear Algebra, 2(1974) 13-24.

J.L. van Hemmen and Ando *An inequality for trace ideals*, Commun. Math. Phys., 76(1980) 143-148.

J. von Neumann, *Some matrix inequalities and metrization of matric space*, Tomsk. Univ. Rev., 1(1937) 286-300, reprinted in *Collected works*, Pergamon Press, 1962.

B. Wang and M. Gong, *Some eigenvalue inequalities for positive semidefinite matrix power products*, Linear Algebra Appl., 184(1993) 249-260.

P.A. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT (13) 217-232.

H.F. Weinberger, *Remarks on the preceding paper of Lax*, Comm. Pure Appl. Math., 11 (1958) 195-196.

H. Weyl, *Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen*, Math. Ann., 71(1911) 441-469.

H. Whitney, *Complex Analytic Varieties*, Addison Wesley, 1972.

H. Wielandt, *Lineare Scharen von Matrizen mit reellen Eigenwerten*, Math. Z., 53(1950) 219-225.

H.W. Wielandt, *An extremum property of sums of eigenvalues*, Proc. Amer. Math. Soc., 6(1955) 106-110.

H.W. Wielandt, *Topics in the Analytic Theory of Matrices*, mimeographed lecture notes, University of Wisconsin, 1967, reprinted in *Collected Works*, Vol. 2, W.de Gruyter, 1996.

E. Wigner and J. von Neumann, *Significance of Löwner's theorem in the quantum theory of collisions*, Ann. of Math., 59 (1954) 418-433.

J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

Index

Notations

Graduate Texts in Mathematics

continued from page ii

119 ROTMAN. An Introduction to Algebraic Topology.
120 ZIEMER. Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.
121 LANG. Cyclotomic Fields I and II. Combined 2nd ed.
122 REMMERT. Theory of Complex Functions. *Readings in Mathematics*
123 EBBINGHAUS/HERMES et al. Numbers. *Readings in Mathematics*
124 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part III.
125 BERENSTEIN/GAY. Complex Variables: An Introduction.
126 BOREL. Linear Algebraic Groups.
127 MASSEY. A Basic Course in Algebraic Topology.
128 RAUCH. Partial Differential Equations.
129 FULTON/HARRIS. Representation Theory: A First Course. *Readings in Mathematics*
130 DODSON/POSTON. Tensor Geometry.
131 LAM. A First Course in Noncommutative Rings.
132 BEARDON. Iteration of Rational Functions.
133 HARRIS. Algebraic Geometry: A First Course.
134 ROMAN. Coding and Information Theory.
135 ROMAN. Advanced Linear Algebra.
136 ADKINS/WEINTRAUB. Algebra: An Approach via Module Theory.
137 AXLER/BOURDON/RAMEY. Harmonic Function Theory.
138 COHEN. A Course in Computational Algebraic Number Theory.
139 BREDON. Topology and Geometry.
140 AUBIN. Optima and Equilibria. An Introduction to Nonlinear Analysis.
141 BECKER/WEISPFENNING/KREDEL. Gröbner Bases. A Computational Approach to Commutative Algebra.
142 LANG. Real and Functional Analysis. 3rd ed.
143 DOOB. Measure Theory
144 DENNIS/FARB. Noncommutative Algebra.
145 VICK. Homology Theory. An Introduction to Algebraic Topology. 2nd ed.
146 BRIDGES. Computability: A Mathematical Sketchbook.
147 ROSENBERG. Algebraic *K*-Theory and Its Applications.
148 ROTMAN. An Introduction to the Theory of Groups. 4th ed.
149 RATCLIFFE. Foundations of Hyperbolic Manifolds.
150 EISENBUD. Commutative Algebra with a View Toward Algebraic Geometry.
151 SILVERMAN. Advanced Topics in the Arithmetic of Elliptic Curves.
152 ZIEGLER. Lectures on Polytopes.
153 FULTON. Algebraic Topology: A First Course.
154 BROWN/PEARCY. An Introduction to Analysis.
155 KASSEL. Quantum Groups.
156 KECHRIS. Classical Descriptive Set Theory.
157 MALLIAVIN. Integration and Probability.
158 ROMAN. Field Theory.
159 CONWAY. Functions of One Complex Variable II.
160 LANG. Differential and Riemannian Manifolds.
161 BORWEIN/ERDÉLYI. Polynomials and Polynomial Inequalities.
162 ALPERIN/BELL. Groups and Representations.
163 DIXON/MORTIMER. Permutation Groups.
164 NATHANSON. Additive Number Theory: The Classical Bases.
165 NATHANSON. Additive Number Theory: Inverse Problems and the Geometry of Sumsets.
166 SHARPE. Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.
167 MORANDI. Field and Galois Theory.
168 EWALD. Combinatorial Convexity and Algebraic Geometry.
169 BHATIA. Matrix Analysis.